THE ROLE OF LONG-TERM MEMORY IN AUTOMATICITY DEVELOPMENT

Rui Cao

Submitted to the faculty of the University Graduate School

In partial fulfillment of the requirements

For the degree

Doctor of Philosophy

In the Department of Psychological and Brain Sciences

Indiana University

August 2018

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____

Robert Nosofsky, PhD

_____

Richard Shiffrin, PhD

_____

Joshua Brown, PhD

_____

Thomas Busey, PhD

August 17, 2018

Acknowledgments

I would like to thank everyone on my dissertation committee for their input on this work: Thomas Busey, Joshua Brown and most importantly, both of my advisers: Richard Shiffrin and Robert Nosofsky. Any graduate student would call him or herself lucky if he or she had opportunity to work with either one of my advisers; but to work with both, it has truly been a privilege. Their insistence on aiming for only the best scientific research will stay with me.

I would also like to thank everyone in Shiffrin lab: Suyog Chandramouli, Sam Harding, and Kiran Kumar for their suggestions to my work, often during our infamously long lab meetings. Graduate school is hard, but it would have been much harder without friends and comrades.

Finally, to my wonderful boyfriend, Arni... I would like to thank you for always believing in me, even when I didn't. Your encouragement and occasional help with last-minute proof-reading from across the Pacific Ocean made this dissertation possible.

Rui Cao

THE ROLE OF LONG-TERM MEMORY IN AUTOMATICITY DEVELOPMENT

Automaticity is extremely common in our daily lives: we perform many routine tasks (e.g. reading) effortlessly with little thought or conscious awareness. In one of the most famous studies published in the field of cognitive psychology, Shiffrin and Schneider (1977) demonstrated how automaticity could be achieved with long training that mapped stimuli to responses consistently (denoted CM). Their demonstrations used visual and memory search for small numbers of items. The many years since those reports notwithstanding, the precise cognitive and neurological mechanisms that underlie the development of automaticity remain elusive. My thesis aims to explore memory search with empirical studies and in particular with quantitative modeling to specify the way that automaticity develops, the rate at which it does so, and the degree to which its development is an automatic consequence of training. To address this issue with computational modeling, I adapted the Exemplar-Based-Random-Walk (EBRW) model. This model has provided excellent accounts of accuracy data and response time data in categorization learning. I extended EBRW to incorporate well-established theories about automaticity learning, specifically, learning of item-response associations in long-term memory. The resultant models were applied to tasks mixing items that were and were not trained consistently, and were compared to alternatives that produced behavior as a consequence of other well-known processes such as decisions based on familiarity. The results demonstrate that the development and use of automaticity is not simply a matter of consistent training, and shows the importance of strategies. A study with measures from an electroencephalogram provided further insights into the processes used to carry out memory search. Both the empirical studies and the

modeling suggest that the development of automaticity is a result of a complex interaction of attention, strategy, memory, and learning.

_____

Robert Nosofsky, PhD

_____

Richard Shiffrin, PhD

_____

Joshua Brown, PhD

_____

Thomas Busey, PhD

Table of Contents

Introduction

Automaticity is extremely common in our daily lives: we perform many routine tasks (e.g. reading) effortlessly with little thought or conscious awareness. In one of the most famous studies published in the field of cognitive psychology, Shiffrin and Schneider (1977) demonstrated how automaticity could be achieved through training using consistent response mapping (CM). They showed this in hybrid memory/visual search tasks. In such tasks, observers search for the presence of one of several to-be-remembered target objects embedded in visual displays. In CM variants of these tasks the mapping of targets and foils to responses remains fixed across all trials. As CM training proceeded, the task became "automatic", characterized with short response time, few errors, and difficulty to reverse the response mapping despite subjects' conscious efforts. The results are in sharp contrast with those using varied mapping (VM), in which targets on some trials are foils on other trials, and vice versa. In VM, the performance improved very little even after extensive training. Shiffrin and Schneider's results established boundary conditions for the development of automaticity and revealed interesting interactions between memory, categorization, and attention. Their studies and others, both before and after, such as those by Posner and Snyder (1975) and by Logan (1988), have had a huge influence on the field, but there remain many questions unanswered. In particular, it remains to produce formal quantitative models of the cognitive and neurological mechanisms that underlie the development of automaticity in short term memory tasks. Understanding these detailed mechanisms will not only lead to theoretical advances in several subfields of psychology and cognitive science, but also has the potential for real world application in skill acquisition. This dissertation therefore carries out empirical studies of the mechanisms that underlie the

development of automaticity in short-term memory studies and focuses on development of formal quantitative modeling of the results.

I start with a conceptual framework proposed in Shiffrin and Schneider (1977): In VM tasks, subjects used a control process requiring capacity in short-term memory. That capacity limit produced performance that depended strongly on the number of items held in memory (the memory set size) and the number of items in the visual display (the visual set size). However, as CM training proceeded, two types of automaticity developed: 1) attention tended to be attracted automatically to any consistent target in the visual display, reducing the visual set size effect, because the first item assessed would usually be the target. 2) the subjects learned the associations (the correct responses) to each of the memory set items, stored these S-R associations in long-term memory, and used their long-term memory to respond, bypassing the capacity limits of short-term memory, and reducing the memory set size effect. It is the second of these processes that is the subject of my thesis. The studies will all present memory items sequentially, the number varying in different trials and conditions, and then test a single item for old-new recognition, in other words whether the test item had been on the study list. The subject is instructed to respond as quickly and accurately as possible, and measures of both accuracy and response time will provide the data used for inference and modeling. Thus the thesis will explore the development of automaticity in these paradigms, and will produce and assess computational models to help understand how learning takes place.

Sternberg (1966) showed that subjects' performance decreased as the memory set size grew. His paradigms allowed time for rehearsal, both during list presentation and prior to test. Much of our recent research has speeded the presentation rate and reduced the time until test, largely eliminating the opportunity for rehearsal. The 'serial-exhaustive' search model Sternberg

developed does not fit the results when rehearsal is greatly reduced (e.g., McElree & Dosher, 1989; Monsell, 1978; Nosofsky, Little, Donkin, & Fific, 2011), and new models have been developed (for a comprehensive review and analysis of set-size effect, see Sternberg, 2016). Nonetheless, VM studies in both versions show large memory set size effects and demonstrate that subjects search the most recent list of presented items and thereby engage capacity limitations that produce these set size effects.

However, quite a different picture emerges when CM training is employed: A series of recent studies (e.g. Cao, Nosofsky, and Shiffrin, 2016; Nosofsky, Cox, Cao and Shiffrin, 2014) showed that the set-size effect is greatly reduced and largely disappears. Furthermore, such improvements occur very early in training. Such results suggest a rapid switch to the use of learned associations in long-term memory, although as will be seen, that is not the only process that can produce such effects. The use of associations stored in long-term memory is not restricted to CM: A recent study (Nosofsky, Cao, Cox and Shiffrin, 2014) has shown such use even in VM training. The study revealed that subjects' performance is affected by item-response associations stored during lists studied and tested prior to the current list.

These complex patterns of results raise questions concerning the way that information from both short-term and long-term memory is combined to carry out memory search, the way the processes differ in VM and CM paradigms, and the role of different strategies in such tasks. Thus one core set of issues I address in my dissertation is the importance and effect of strategies.

To address the issues that have been raised, I use computational modeling applied to the varied results. In particular, I adapted the Exemplar-Based-Random-Walk (EBRW) model developed by Nosofsky and Palmeri (1997) to account for speeded classification. This model has provided excellent accounts of accuracy data and response time data from both categorization

tasks and short-term probe-recognition memory tasks (Nosofsky et al. 2011). To deal with results showing that prior lists affect performance in VM, and the effects of long-term learning in CM, I extended EBRW to incorporate well-established theories about automaticity learning, specifically, learning of item-response associations. The resultant model is termed the "IR-model". I contrast this model with an alternative that assumes a familiarity-only process, with little or no learning, termed the "FAM-model". To explain how and when FAM- and IR-models are used in these tasks, I describe how training and strategy affect their use and affect performance.

In Chapter 1, I present a study manipulating the degree of CM training or VM training for individual items. The study addresses a concern about global or local factors that might produce automaticity: If CM training mixes items with different frequencies, will global automaticity develop for all items, and will the degree of automaticity be equal for all CM items, or perhaps be governed by each item's degree of training? The behavior pattern and the computational modeling results showed that the IR model was able to provide a good account for the CM condition by assuming learning at individual item level. Moreover, contrary to what is commonly assumed, the data pattern in the VM condition was better captured by the IR model than the FAM model. These results suggest that item-response associations might be stored and used quite generally in early stages of learning, thereby affecting performance in both CM and VM.

In Chapter 2, I further tested the possibility that item-response learning plays a general role in memory search tasks, for both CM and VM tasks. Two experiments mixed CM and VM items within a single trial during training. The first mixed VM and CM items. The second mixed VM and CM items with new items not yet experienced, all within trials. The within-trial mixing

makes it likely that the subjects use the same strategy for all items. Will the IR model fit this study, or will subjects abandon item-response learning or at least the use of such learning to perform, since in most situations including the present one it is suboptimal to try to perform using item-response associations for VM and new items in a study list. The results from both experiments challenged the idea of item-response learning as the unified mechanism and suggested that subjects were very sensitive to the specific training conditions.

In the first two chapters I focused mostly on the way that retrieval operates in short-term memory tasks, and how this varies with the development of automaticity (in the form of IR learning and use). In Chapter 3, I focus specifically on the study phase, the period when subjects encode the presented items. I use neural measurements (EEG measures) as each item is studied to examine the different way that items are encoded in VM and CM tasks. Shiffrin and Schneider (1977) proposed that attention allocated to encode information in each trial was largely reduced as CM training proceeded. This idea was confirmed in previous studies by Woodman and colleagues (e.g. Carlisle, Arita, Pardo & Woodman, 2011; Woodman, Carlisle & Reinhart, 2013) using simple memory items presented together in a single display. In Chapter 3 I report a new experiment that used our standard paradigm, with sequential presentation and with complex stimuli. We trained subjects in CM and VM conditions in separate blocks. The EEG findings and the behavioral results confirmed the idea that CM training reduced the need to rely on limited capacity short-term memory, but also raised additional questions about the degree to which EEG measures reflect short-term memory load.

The results from these studies and the modeling applied to the results suggest that the development of automaticity is a result of an intricate interaction of attention, strategy and

memory systems. A model that aims to provide a comprehensive account of the development of automaticity will need to incorporate all these elements[1].

Footnotes

1. The first chapter and the third chapter in the dissertation are published manuscripts (Cao, Shiffrin, & Nosofsky, 2018; Cao, Busey, Nosofsky, Shiffrin, & Woodman, 2018). The second chapter is a manuscript for future publication.

References

Cao, R., Busey, T.A., Nosofsky, R. M., Shiffrin, R. M., & Woodman, G.F. (2018). Tracking the Development of Automaticity in Memory Search with Human Electrophysiology. *40th Annual Cognitive Science Society Meeting Proceedings*.

Cao, R., Nosofsky, R. M., & Shiffrin, R. M. (2017). The development of automaticity in short-term memory search: Item-response learning and category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43(5),* 669.

Cao, R., Shiffrin, R. M., & Nosofsky, R. M. (2018). Item frequency in probe-recognition memory search: Converging evidence for a role of item-response learning. *Memory & Cognition*

Carlisle, N. B., Arita, J. T., Pardo, D., & Woodman, G. F. (2011). Attentional templates in visual working memory. *Journal of Neuroscience, 31, 9315–9322*.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*, 492–527.

McElree, B., & Dosher, B. A. (1989). Serial position and set size in short-term memory: Time course of recognition. *Journal of Experimental Psychology: General, 18*, 346–373.

Monsell, S. (1978). Recency, immediate recognition memory, and reaction time. *Cognitive Psychology, 10*, 465–501.

Nosofsky, R. M., Cao, R., Cox, G. E., & Shiffrin R. M. (2014). Familiarity and categorization processes in memory search. *Cognitive Psychology, 75*, 97-129.

Nosofsky, R. M., Cox, G. E., Cao, R., & Shiffrin, R. M. (2014). An exemplar-familiarity model predicts short-term and long-term probe recognition across diverse forms of memory search. *Journal of Experimental Psychology: Learning, Memory, and*

Cognition, *40*(6), 1524.

Nosofsky, R.M., Cao, R., Cox, G.E., & Shiffrin, R.M. (2014). Familiarity and categorization processes in memory search. Cognitive Psychology, 75, 97-129.

Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review, 188*, 280–315

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104(2), 266-300*.

Posner, M.I., & Snyder, C. R. R. (1975). Attention and cognitive control. In R. I., Solso (E.d.), *Information processing and cognition: The Loyola Symposium* (pp. 55-85). Hillsdale, NJ: Erlbaum.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84*, 127–190.

Sternberg, S. (1966, August 5). High-speed scanning in human memory. *Science, 153*, 652– 654. http://dx.doi.org/10.1126/science.153.3736.652

Sternberg, S. (2016). In defence of high-speed memory scanning. *The Quarterly Journal of Experimental Psychology*, *69*(10), 2020-2075.

Reinhart, R. M. G., Carlisle, N. B., & Woodman, G. F. (2014). Visual working memory gives up attentional control early in learning: Ruling out inter-hemispheric cancellation. *Psychophysiology, 51*(800-804).

Item frequency in probe-recognition memory search: Converging evidence for a role of item-response learning

Probe-recognition memory-search tasks are among the most common paradigms for studying memory. In these tasks, a list of to-be-remembered items (the "memory set") is presented, followed by a test probe that either is a member of the memory set (an "old" probe or "target") or not a member of the memory set (a "new" probe or "foil"). Subjects aim to make the old/new judgment as quickly and accurately as possible. Both accuracy and response time (RT) are recorded to measure subjects' performance. As observed in the original Sternberg (1966) studies, RT increases substantially as the size of the memory set increases, a result termed the *set-size* effect. The detailed processes that operate when participants engage in the memory-search task may vary with details of the experimental conditions (e.g., McElree & Dosher, 1989; Monsell, 1978; Nosofsky, Little, Donkin, & Fific, 2011; Sternberg, 1966; for a comprehensive review and analysis, see Sternberg, 2016). For present purposes, however, the key point is that the presence of the memory set-size effect provides a clear indication that the observers' engagement with the current set in short-term memory plays a fundamental role in determining performance.

In their studies that examined hybrid forms of memory/visual search, Schneider and Shiffrin (1977) discovered that the set-size effect could be greatly reduced or eliminated under "consistent mapping" (CM) conditions. In CM, the old (target) probes are chosen from one fixed set of items on every trial (termed the "positive set"), and the new (foil) probes are chosen from a separate fixed set of items on every trial (termed the "negative set"). Thus, the old targets and the new foils never switch roles across trials. As practice proceeded in Schneider and Shiffrin's

studies, performance improved dramatically: subjects were able to make their old/new judgments with shorter RT and fewer errors. Most importantly, the performance became largely invariant to the set-size manipulation, suggesting reliance on a process other than the retrieval of the list held in short-term memory (see also Logan & Stadler, 1991). In Schneider and Shiffrin (1977), subjects were also tested in a varied-mapping (VM) condition, where the items that served as old probes on some trials were new probes on other trials, and vice versa. In contrast to CM, performance in the VM condition improved very little with practice, and the set-size effect persisted even after extensive practice. The researchers proposed that performance in the VM condition required an effortful, controlled process, regardless of the amount of practice; whereas practice in the CM condition allowed for the development of an extremely efficient, automatic form of information processing.

Although Shiffrin and Schneider (1977) developed a conceptual framework for understanding the nature of the controlled and automatic processes that developed in these tasks, their modeling did not delve deeply into the quantitative details of the processes at work. One influential model that aims to provide a quantitative account of the development of automaticity is Logan's (1988) instance theory. According to instance theory, highly efficient automatic performance arises by retrieving responses that are linked to the instances stored in long-term memory. With increased practice, more instances are stored in memory, which leads to a more efficient retrieval process (see Logan, 1988, 1990 for details). However, instance theory focuses only on how behavior changes in CM training and therefore does not provide a detailed account for the difference between CM and VM performance. Strayer and Kramer (1994) developed some descriptive accounts based on diffusion modeling (Ratcliff, 1978) to characterize the differences across CM and VM data patterns. They concluded that the difference reflects changes

in both drift rates (i.e., rates of evidence accumulation) and response thresholds. Although the researchers further interpreted the diffusion-model parameters from the perspective of strategic vs. learning factors, the aim of the paper was not to develop a mechanistic account of the cognitive processes that give rise to the different evidence accumulation rates in the CM and VM conditions. The main goal of the current work is to fill that gap and move toward the development of a process-level model that provides a quantitative account of performance in both CM and VM memory-search tasks. Because it is often difficult to derive precise predictions from theories that are specified at a purely verbal level, and because results from VM and CM memory-search tasks have been extremely influential in guiding thinking about the development of automacity, this goal of developing a formal mathematical process-model for VM and CM memory search is clearly a highly significant one.

Some progress towards this goal was made in recent work by Nosofsky, Cao, Cox, and Shiffrin (2014; Nosofsky, Cox, Cao, and Shiffrin, 2014). The formal model is an extended version of the exemplar-based random-walk (EBRW) model that has been successfully applied to various forms of categorization (Nosofsky & Palmeri, 1997; Nosofsky & Stanton, 2005) and old/new recognition memory (Nosofsky, et al., 2011). In the version of the model applied to probe-recognition memory search, each item of the memory set is stored as an exemplar in short-term memory. Items from previous memory sets may also be stored in long-term memory. When the test probe is presented, it activates exemplars to which it is similar (both short-term and long-term), and the activated exemplars race to be retrieved (see Formal Modeling section for details). In an "item-familiarity-only" version of the model, each retrieved exemplar leads an information accumulator to move toward an "old" threshold; while failure to retrieve an old exemplar leads the information accumulator to move toward a "new" threshold. The retrieval

process continues until one or the other response threshold is reached, at which time the observer emits the response that is associated with that threshold.

In initial tests, Nosofsky, Cox et al. (2014) found that, with appropriate parameter settings, this item-familiarity version of the model provided excellent accounts of both VM and CM accuracy and RT patterns across conditions with a wide range of list lengths (memory set sizes of 1, 2, 4, 8 and 16). However, in subsequent work, Nosofsky, Cao et al. (2014) obtained evidence that clearly challenged the item-familiarity-only account of CM performance. In particular, in this study, the researchers examined cases in which the same stimulus served as a test probe across two consecutive trials. A key result was that in VM, there was massive interference in responding "new" to "new" test probes if that probe had been presented on the previous trial (for similar previous findings, see, e.g., Monsell, 1978). Crucially, however, there was no such interference in CM: if anything, repeating a new test probe across two consecutive trials led to slight facilitation.

Nosofsky, Cao et al. (2014) suggested that in VM, the observer relied on an item-familiarity process: Recent past presentations of items boost their familiarity on the current trial, leading to greater tendencies to respond "old" to such items. Thus, one would observe interference across trials in which a new test probe was repeated. By contrast, the researchers interpreted the facilitation observed in the repeated trials of the CM condition as evidence that observers instead relied on remembered *item-response* mappings in that condition. In particular, the idea is that the old-new response associated with each test probe is stored along with that test probe on each trial of the experiment. In later trials, retrieval of exemplars with "old" response labels would drive the random walk toward the old response threshold, but retrieval of exemplars with "new" response labels would drive the random walk toward the "new" response threshold.

Thus, in the CM condition, when a new test probe is repeated across trials, the item-response learning that took place would facilitate the "new' response on the current trial. Nosofsky, Cao et al. (2014) developed a formal model to implement these ideas and it yielded good quantitative accounts of the full range of individual-subject performance across more than 30 sessions of CM and VM practice.

Although the model performed well, the crucial empirical result that motivated the model was derived from cases in which test probes repeated across consecutive trials. A concern that arises is that the observer might have access to the item-response label only for exemplars that have been presented very recently. In other words, the facilitation observed in CM for the repeated new stimuli may be a byproduct of more vivid memory traces from very recent presentations rather than arising from more durable long-term associations. Such a view is consistent with the finding that the lag with which items are presented on current lists often exerts a powerful effect on short-term probe recognition (e.g., McElree & Dosher, 1989; Nosofsky et al., 2011). In addition, participants may apply a special-purpose strategy to take advantage of the repeated-trials manipulation, but the strategy may have little generality across more usual conditions of CM training.

The present study addressed these concerns by varying the long-term frequency with which individual items were presented in both the CM and VM conditions. Clearly, the more frequently presented items would give rise to higher long-term familiarity. Thus, to the extent that item-familiarity mechanisms play the sole role, one would expect the high-frequency items to lead to greater tendencies to respond "old" (for both old and new test probes) in both the VM and CM conditions. Thus, one should observe *interference* effects for high-frequency new test probes in both VM and CM. By contrast, suppose instead that observers form long-term *item-*

response associations in CM. Increasing the frequency of the consistent pairings should boost the strength of those item-response associations. Thus, even for the new test probes in the CM condition, one should see *facilitation* in performance for the high-frequency items (compared to the low-frequency ones), in direct contrast to the predictions from the item-familarity model.

Finally, although most past accounts of the influence of LTM on VM recognition-memory performance involve familiarity-only mechanisms, we were also interested in exploring the extent to which item-response-learning mechanisms might play some role in VM as well. As will be seen, models that formalize the role of these item-familiarity vs. item-response-learning factors also make very different predictions concerning the patterns of results that will be observed when presentation frequency is manipulated in the VM condition. We defer the precise statement of these predictions until after presentation of the formal model that guides the research.

Experiment

We tested subjects in both VM and CM probe-recognition memory-search tasks. In both tasks, we manipulated memory-set size. The key manipulation was to also vary the frequency with which individual items were presented across trials in both the VM and CM tasks.

Method

Subjects

The subjects were 109 undergraduate students from Indiana University who participated in partial fulfillment of an introductory psychology course requirement. Subjects were randomly assigned to either the CM condition (55 subjects) or the VM condition (54 subjects).

Stimuli and Apparatus

The stimuli were drawn from a pool of 2,400 unique object images used and described by Bradly, Konkle, Alvarez, and Oliva (2008). Each image subtended a visual angle of approximately 7 degrees and was displayed in the center of a gray background. The experiment was conducted on PCs using MATLAB and the Psychophysics Toolbox (Brainard, 1997). All subjects were tested individually in sound-attenuating cubicles.

Procedure

In all conditions, half the test probes were targets and half were foils, with type of test probe chosen randomly on each trial. The memory-set sizes were 2, 4 and 6; memory set size was chosen randomly on each trial. For each subject, 32 stimuli were randomly sampled from the 2,400 images. On each trial in the VM condition, the memory set was randomly sampled from the 32-stimulus set, subject to the constraints of an item-frequency manipulation described below. Targets were randomly chosen from the memory set; foils were randomly chosen from the remaining items in the 32-item set. In the CM condition, for each subject,16 stimuli were randomly drawn from the 32-stimulus set and served as the "positive set" on all trials; the remaining 16 stimuli served as the "negative set" on all trials. On each trial, the memory set was randomly sampled from the positive set, subject to the constraints of the item-frequency

manipulation (see below). Target test probes were randomly chosen from the memory set; foil test probes were randomly selected from the negative set.

Item frequency was manipulated as follows. In both the VM and CM conditions, items in each subject's stimulus set were randomly divided into high frequency (HF), medium frequency (MF) and low frequency (LF) roles. HF items were assigned a "selection weight" of 10; MF items a selection weight of 5; and LF items a selection weight of 1. For each subject, the CM positive set contained 2 HF items, 2 MF items and 12 LF items. The following sequential-selection algorithm was used for constructing the memory set on each trial of the CM condition. Let w_i denote the selection weight associated with item i. Then the probability that item i was the first item selected was given by $w_i / \sum w_k$. Next, the probability that item j was the second item selected (from among the remaining positive-set items) was given by $w_j / \sum_{k \neq i} w_k$, where $k \neq i$ denotes that the sum is across all items not including i. The item selections continued in analogous fashion until the memory set size was reached. (Note that although the memory-set items were selected using the just-described sequential algorithm, the serial positions of the selected items in the presentation sequence were chosen at random.) For target trials, the test probe was randomly drawn from the memory set; each memory-set item had an equal probability of serving as the test probe regardless of the assigned weights. The CM negative set had the same structure as the CM positive set (i.e., 2 HF items, 4 MF items and 12 LF items). Test items that were foils were selected from the CM negative set with probability equal to their relative selection weights (i.e., $w_i / \sum w_k$).

The item-frequency manipulation in the VM condition was analogous to the one just described for the CM condition. For each subject, the VM set contained 4 HF items, 4 MF items and 24 LF items, with selection weights as described above. The probability that item i was the

first item selected for inclusion in the memory set was given by $w_i / \sum w_k$; the probability that item j was then the second item selected was given by $w_j / \sum_{k \neq i} w_k$; and so forth until the memory set size was reached. For target trials, the test probe was randomly drawn from the selected memory set; each memory-set item had an equal probability of serving as the test probe. For foil trials, the test probe was randomly selected from among those items not in the memory set, with probability proportional to its assigned selection weight.

The relative proportion of trials with which the different item types served as memory-set items, old test probes, and new test probes in the actual experiment is reported in Table 1. Inspection of the table confirms that, in both the CM and VM conditions, and at both test and study, individual HF items occurred with the highest frequency, followed by individual MF items and finally individual LF items. In addition, the total probability with which each of the individual item types appeared as test probes was roughly equated across the CM and VM conditions. Of course the relative frequency with which individual item types were assigned to specific responses differed across CM and VM: for example, in CM an HF item from the positive set would always appear as an old test probe; in VM an HF item would appear roughly half the time as an old test probe and half the time as a new test probe.

Each trial began with the presentation of a fixation point (asterisk) in the center of the screen for 0.1 second, followed by the presentation of the memory set. Each memory set item was presented in the center of the screen for 1 sec with a 0.1 sec inter-stimulus interval. After a 1 sec retention interval, a second fixation point (plus sign) was presented for 0.5 sec, followed by the presentation of the test probe. The test probe remained on the screen until subjects responded (by pressing the 'F' or 'J' key on the computer keyboard). Feedback ("Correct!" or "Incorrect") was then provided for 1 sec. Each subject completed 5 blocks of testing with 25 trials per block.

The computer reported to the subjects their overall percentage of correct responses at the end of each block. Each block took approximately 5 minutes to complete, with the entire session lasting approximately 30 minutes. Subjects were instructed to make their responses as quickly and accurately as possible. Subjects were not alerted to the possibility that some items would repeat frequently across trials; nor were they alerted to the differing structures of the VM versus CM conditions.

Results

We considered the first block to be a practice block, so did not include the data from the first block in our analyses. Although our inclusion of both MF and HF items was originally intended to yield stronger parametric constraints for model fitting, inspection of the data indicated similar results for the HF and MF items. To reduce noise in the data, we combined the results from the HF and MF trials (and refer to both as "HF"). (Combining HF and MF can be theoretically justified by assuming that strength in memory increases in negatively accelerated fashion with item repetition, so the MF items are much closer in strength to the HF items than to the LF ones.) Also, because our investigation was intended to investigate the effects of long-term frequency on CM and VM performance, we considered trials in which the same test probe was repeated from an immediately previous trial to be a special case and removed the few such trials from analysis (~0.16% trials). Trials with response time (RT) greater than 5000 ms or less than 180ms were also eliminated (~0.7% trials). We then calculated the mean and standard deviation of RT for each Condition (CM vs. VM) x Set Size x Probe Type (old vs. new) x Frequency (HF vs. LF) x Lag combination and discarded trials that were greater than 2.5 standard deviations away from the mean (~3% trials). Finally, we eliminated the data from three outlier subjects in the CM condition who performed significantly worse than the remaining

subjects in the group (overall median RT greater than 1500 ms or overall proportion correct less than 0.8); and eliminated the data of three outlier subjects in the more difficult VM condition (overall median RT greater than 2000 ms or overall proportion correct less than 0.6).

The main results of the experiment are displayed in the left panels of Figures 1 and 2. In Figure 1 we plot the mean RTs for correct responses as a function of tasks (CM vs. VM), set size, type of test probe (old vs. new) and item frequency. The error probabilities are plotted as a function of these variables in Figure 2. The error bars indicate between-subjects standard errors. Ignoring for a moment the effects of the item-frequency manipulation, the overall data pattern is highly consistent with that of recent studies using a similar paradigm and set of materials (e.g. Nosofsky, Cox et al. 2014): Performance in the CM condition (top row of each figure) is better than in the VM condition (bottom row of each figure), with both lower error rates and shorter RTs. Moreover, there is little if any effect of set size for new items in the CM condition, but a big effect of set size for new items in the VM condition. Both conditions show set-size effects for old items, although the effects tend to be smaller in the CM condition than in the VM condition. A more detailed breakdown of the old-item data is shown in Figures 3 and 4, which plots performance on the old items as a joint function of set size and lag of presentation (where lag is measured backwards from the end of the study list). Although the plots are noisy due to small sample sizes, they basically replicate patterns we have observed in closely related experiments (Nosofsky et al., 2011, 2014a,b): overall performance on old items gets worse with increases in lag, with little if any additional effect of set size once one conditions on lag. The main basis for the set-size dependence seen in Figures 1 and 2 is the fact that larger set sizes include items with longer lags.

The key new results of interest involve the effects of the item-frequency manipulation. As can be seen in Figures 1 and 2, in the CM condition, overall performance tends to be better for the HF items than for the LF items, although the locus of the effect varies across test-probe types and performance measures. Specifically, for the new test probes, mean RTs are shorter for HF than LF (with little difference in error probability, which is near floor for both HF and LF). For the old test probes, error probability is lower for HF than for LF (with little difference in the mean RTs). Because the locus of the effect differs for the old and new probes, we conducted separate statistical analyses for them. We analyzed the CM data using a 3 (set size: 2 vs 4 vs 6) x 2 (HF vs LF) repeated-measures ANOVA. The effect of the frequency manipulation on mean correct RTs for the new test probes was significant, $F(1,51) = 12.96$, $p< 0.001$. In addition, overall mean correct RTs for the old HF probes were significantly shorter than for the old LF probes, $F(1, 51) =4.59$, $p = 0.037$ (although the effect appears to be restricted to only the largest set size).[1] The effect of the frequency manipulation on error probability for the old test probes was also significant: $F(1, 51) = 8.26$, $p = 0.006$. There is no evidence for an effect of the frequency manipulation on error probability for the new test probes, $F(1, 51)<1$; however, error probability is near floor for both HF and LF, so the lack of an effect is not surprising.

In sum, combining the patterns of accuracy and correct mean-RT data, overall performance on both the old *and* new test probes is better for the HF items than the LF items in the CM condition. These results are consistent with the hypothesis that item-response-learning governed performance in the CM condition: For both old and new items, performance is benefited by increases in the frequency of consistent item-to-response training. The results are inconsistent with the hypothesis of a pure item-familiarity hypothesis for the CM condition. HF new items are far more familiar than are LF new items. According to the item-familiarity

hypothesis, increased familiarity should lead to increases in "old" responding. For new items in the CM condition, however, the RT results point decidedly in the opposite direction.

Our results for CM are reminiscent of results from a hybrid memory/visual-search paradigm reported recently by Wolfe, Boettcher, Josephs, Cunningham and Drew (2015). In these studies, subjects repeatedly searched visual displays for the presence of targets from a single memorized list. The key manipulation across experiments was to vary the familiarity of foils that appeared in the visual displays. The general finding was that foil familiarity exerted little if any impact on visual-search performance (either in terms of false-alarm rates or slowed RTs), leading Wolfe et al. to conclude that, under CM conditions, item-familiarity mechanisms do not cause observers to confuse foils with targets. (Wolfe et al. did not test VM versions of their task.) Our present results for CM in pure memory-search tasks converge with those observed by Wolfe et al. in their hybrid memory-visual search tasks. Indeed, our results suggest that increased foil frequency can *benefit* the process of rejecting CM foils.

In direct contrast to the CM task, in the VM task overall performance is *worse* for the HF items than for the LF items (see Figures 1 and 2). Furthermore, this results holds for both performance measures (error probabilities and RTs) for both new *and* old probes. To analyze the data, we conducted a 2 (test-probe type: old vs new) x 3 (set size: 2 vs 4 vs 6) x 2 (frequency: HF vs LF) repeated measures ANOVA. The analysis yielded a significant main effect of item frequency on both error probability ($F(1,50) = 18.66$, $p<0.001$) and correct RT ($F(1,50) = 8.75$, $p=0.005$), reflecting the worse overall performance associated with the HF items. There was also a significant interaction between test-probe type and frequency ($F(1, 50) = 11.74$, $p = .001$ for the error data; $F(1, 50) = 11.54$, $p = .001$ for the RTs). The interaction reflects the finding that

whereas there was a big effect of item frequency for the new probes, there was only a trend for the old probes.

In the VM task, it is not surprising that HF new items are classified more slowly and with greater error probabilities than are LF new items. Such an effect is predicted by the item-familiarity hypothesis of VM performance. In particular, HF new items will tend to have far greater long-term familiarity than LF new items, which should interfere with observers' ability to classify such test items as "new". More interesting is that there was a trend for the HF *old* items to show a performance deficit compared to LF old items. This pattern of results is the opposite of what would be predicted by a simple item-familiarity hypothesis. Because HF old items have greater familiarity than LF old items, observers should show performance benefits in classifying them as "old", but the results point in the opposite direction.

As we demonstrate in our ensuing Theoretical Analysis section, the overall pattern of results is instead consistent with the idea that an item response-learning mechanism operates not only in the CM condition, but may operate to some extent in the VM condition as well. The key factor is that high-frequency VM items have served as both old and new test probes in numerous previous test trials. There are several ways to implement a mechanism by which this factor could cause interference. The specific approach that we follow is to implement a mechanism that leads the high-frequency inconsistent mappings to result in lowered evidence-accumulation rates in the EBRW memory-search model, resulting in lowered accuracy and longer RTs for the HF items.

Theoretical Analysis

The Formal Models

A schematic illustration of the main components of the EBRW memory-search model[2] is presented in Figure 5. We start by describing the components that are sensitive to the contents of the current study list ("short-term memory"). Then, we expand our description to include contributions from long-term memory as well.

Short-Term Memory Components. According to the model, each of the study items from the current list is stored as an individual exemplar in memory. The memory strength of each exemplar decreases with the lag with which it was presented on the study list. (Lag is measured backwards from the end of the study list.) More specifically, based on evidence reported by Donkin and Nosofsky (2012a; see also Anderson & Schooler, 1991; Wixted & Ebbesen, 1991), it is assumed that memory strength decreases as a power function of lag *j*:

$$m_j = \alpha + j^{-\beta}, \qquad (1)$$

where α is asymptotic strength and β describes the rate of decrease in strength with lag. The differential memory strengths are represented schematically in Figure 5A, where the larger circles represent exemplars with greater memory strength.

When the test probe is presented, the exemplars stored in memory are "activated" and "race" to be retrieved, with rates that are proportional to their activations (cf. Logan, 1988) – see Figure 5B. The degree to which exemplar j (e_j) is activated is a joint function of exemplar j's memory strength and its similarity (*s*) to test-probe i (t_i):

$$a_{ij} = m_j, \quad \text{if } t_i = e_j \qquad (2a)$$

$$a_{ij} = m_j s, \quad \text{if } t_i \neq e_j \qquad (2b)$$

where s ($0 < s < 1$) is a freely estimated similarity parameter. Thus, the study-list exemplars that are most highly activated are those that match the test probe and that have short lags.

As explained in detail in previous articles (e.g., Nosofsky et al., 2011; Nosofsky, Cao et al., 2014), the EBRW-recognition model presumes that the observer establishes "criterion elements" in the memory system. Upon presentation of the test probe, the criterion elements (labeled "c" in Figure 5B) race to be retrieved (along with the stored exemplars). The criterion elements race at a constant rate k, independent of the specific test probe that is presented.

Finally, the retrieved exemplars and criterion elements drive a random-walk process that leads to "Old" versus "New" decisions (Figure 5C). The observer sets response boundaries *+OLD* and *-NEW* that establish the amount of evidence needed for making an "old" or a "new" response. On each step of the random-walk process, if an old exemplar is retrieved, a random-walk counter takes a step toward the "Old" response boundary; whereas if a criterion element is retrieved, the random-walk counter steps toward the "New" response boundary. The retrieval process continues until one of the response boundaries is reached, at which point the observer emits the appropriate response.

Given further technical assumptions concerning the distribution of exemplar race times (see Nosofsky & Palmeri, 1997, p. 268), it turns out that, on each step of the random walk, the probability that the random-walk counter steps toward the *+OLD* response boundary (p_i) is given by:

$$p_i = A_i / (A_i + k), \qquad (3)$$

where A_i is the summed activation of the test probe to all the study-list items:

$$A_i = \sum a_{ij} \qquad (4)$$

and k is the level of criterion-element activation. (The probability that the random walk steps toward the new boundary is simply $q_i = 1 - p_i$.)

Through experience in the task, the observer is presumed to learn an appropriate setting of criterion-element activation k, such that the summed activation (A_i) tends to exceed k when the test probe is old, but tends to be less than k when the test probe is new. Because A_i tends to increase with set size (for "new" test probes), we presume that the observer may adjust the criterion-element activation with changes in set size. As an approximation to implementing possible criterion adjustment, it is assumed that the criterion setting varies linearly with memory set-size M:

$$k(M) = u + v \cdot M. \qquad (5)$$

Long-Term Memory Components. Recent extensions of the EBRW memory-search model (Nosofsky, 2016; Nosofsky, Cao et al., 2014) implement the influence of previous study-test trials (beyond the current study list) with a set of long-term memory (LTM) components (Figure 5B). Specifically, study and test items from the previous trials are presumed to be stored as exemplars in LTM, and race to be retrieved along with the current study-list exemplars. We distinguish between two processes that may mediate the influence of the retrieved LTM exemplars. First, retrievals of LTM exemplars may always drive the random walk towards the +*OLD* response boundary, regardless of the exemplars' status as "old" versus "new" test probes on the previous trials. We denote such a process as an "item-familiarity" (FAM) model and formalize the model with a set of *FAM* parameters. Alternatively, the observer may store along

with the previously tested items their associated "old" versus "new" response labels and retrieve item-response pairs. Retrieval of exemplars with "old" response labels would drive the random-walk towards the *OLD* response boundary, whereas retrieval of exemplars with "new" response labels would drive the random walk towards the *NEW* response boundary. We denote such a process as an "item-response-learning" (IR) model and formalize it with a set of *IR* parameters. The details of both models are described below.

LTM-FAM. In the FAM model, we presume that the activation and retrieval of LTM exemplars always drives the random-walk counter towards the +*OLD* boundary. For simplicity, we account for the boost in the summed activation (A_i) with a free parameter *FAM:*

$$p_i = (A_i + FAM) / [(A_i + FAM) + k]. \quad (6)$$

It is natural to assume that HF test probes receive a greater familiarity boost than do LF test probes. Therefore we estimate separate *FAM* parameters for the HF items versus the LF items (with the constraint that $FAM_{HF} \geq FAM_{LF}$). As discussed in more detail in the model-fitting section, although the model supposes that the same basic process applies across the CM and VM conditions, we allow the *FAM* parameter values to vary across these conditions.

LTM-IR. In the IR model, we presume that the retrieved "item-plus-response-label" exemplars direct the random-walk counter to the response threshold that corresponds with the stored response label. We denote by *IR-OLD* the boost toward the +*OLD* boundary and by *IR-NEW* the boost toward the –*NEW* boundary. Given the structure of the CM condition, old test probes will activate many exemplars with "old" response-labels but no exemplars with "new"

response-labels. Thus, in the CM condition, on trials in which old test probes are presented, the probability that the random walk steps toward *+OLD* is given by:

$$p_i(\text{old}) = (A_i + IR\text{-}OLD) / [(A_i + IR\text{-}OLD) + k]. \quad (7a)$$

Analogously, because new test probes will retrieve only exemplars with "new" response labels, the probability that the random walk steps toward the *–NEW* boundary (if tested with a "new" probe) is given by

$$q_i(\text{new}) = (k + IR\text{-}NEW) / [(k + IR\text{-}NEW) + A_i]. \quad (7b)$$

As is the case for the FAM model, we presume that HF items have strengths in LTM at least as great as LF items, so we introduce the parameter constraints that $IR\text{-}OLD_{HF} \geq IR\text{-}OLD_{LF}$ and that $IR\text{-}NEW_{HF} \geq IR\text{-}NEW_{LF}$. Thus, from inspection of Equations 7a and 7b, it can be seen that for CM the IR model predicts increased evidence-accumulation rates to the correct response boundaries with increases in the frequency of the consistent response mappings.

Unlike in the CM condition, in the VM condition a test probe will activate previous-trial exemplars with both "old" response labels and "new" response-labels, regardless if it serves as an old or a new test probe on the current trial. (The reason is that in VM each item serves randomly as an old test probe and as a new test probe throughout the experiment.) Therefore, for both old and new test probes, the probability that the random walks steps toward the *+OLD* boundary is given by

$$p_i = (A_i + IR\text{-}OLD) / [(A_i + IR\text{-}OLD) + (k + IR\text{-}NEW)]. \quad (8)$$

As was the case for CM, we again presume that $IR\text{-}OLD_{HF} \geq IR\text{-}OLD_{LF}$ and that $IR\text{-}NEW_{HF} \geq IR\text{-}NEW_{LF}$. Thus, from inspection of Equation 8, it can be seen that for VM, the IR model predicts *decreased* rates of evidence accumulation to the correct response boundaries with increases in item frequency.

We should note that if the *IR-OLD* and *IR-NEW* parameters grow indefinitely with frequency and training, then they would come to dominate VM responding, and the current list would not even matter. In a subsequent discussion of the model-fitting results, we provide reasons why the IR parameters are *not* expected to grow indefinitely in VM; this subsequent discussion will also explain why the magnitude of the IR parameters is expected to be lower in VM than in CM.

The full version of the FAM model makes use of 11 free parameters for fitting the data of each condition (for a listing, see Table 3): the lag-related memory-strength parameters α and β (Equation 1); criterion-element parameters *u* and *v*; similarity parameter *s*; response-boundary parameters *+OLD* and *–NEW*; a scaling parameter κ that measures the time of each step in the random walk; a residual-time parameter T_R that reflects non-decision-time processes; and the LTM parameters FAM_{HF} and FAM_{LF}. The IR model has 13 free parameters: the same ones as the FAM model, except instead of using the familiarity-based LTM parameters, it uses the set of item-response LTM parameters: $IR\text{-}OLD_{HF}$, $IR\text{-}OLD_{LF}$, $IR\text{-}NEW_{HF}$, and $IR\text{-}NEW_{LF}$.

Fits of the Models to the Group Data

Because our main goal was to assess the extent to which the alternative models could account for broad, qualitative aspects of the data, we fitted both the FAM and IR models to the averaged group data by minimizing a weighted sum-of-squared deviations (WSSD) criterion. In particular, we required the models to simultaneously fit the mean-correct RT data and the error proportions data of: (a) the new items as a function of set size and (b) the old items as a joint function of set size and lag. To jointly fit all these data sources, we need to apply different weights to the data points (because they are measured on different scales and based on differing sample sizes). We found that a good overall match to both the RT and accuracy data was achieved by minimizing the WSSD with the deviations from the accuracy data (measured in proportions) given twice the weight of the deviations from the RT data (measured in seconds); and the individual data points for new probes given 4 times the weight of the individual data points for the old probes. (Sample sizes for the new-item data points are much greater than for the old-item data points because they are not broken down by lag.)

Based on the theoretical considerations that we described earlier in the *Formal Models* section, we constrained the LTM parameters such that the boosts for the HF items were at least as great as for the LF items (for both the FAM and the IR models). Before imposing any other constraints, we started by fitting the "full" version of both models to the data, with all parameters allowed to vary freely across the VM and CM conditions. The fits of the full models provide baselines for comparison with more constrained versions of the models that we examine subsequently. Because different processes may mediate performance across the CM and VM conditions, we reasoned that it was important to get started by fitting the models separately to the two conditions (i.e,. with all free parameters allowed to vary).

The WSSD fits for different versions of the models are reported in Table 2. The best-fitting parameters from the full version of the FAM model are reported in Table 3 (along with the best-fitting parameters from a constrained version of the IR model that we describe below). Inspection of Table 2 reveals that the WSSD fit for the FAM model is worse than for the IR model for both the CM and VM conditions. Indeed, we will see that even highly constrained versions of the IR model fit the data from both conditions better than does the full version of the FAM model.

To see the reason for the poor fits yielded by the FAM model, we display its predictions of the mean-correct RTs and error probabilities in Figures 1B and 2B, in the same fashion as for the observed data. As can be seen, the FAM model displays various qualitative shortcomings. First, it failed to predict any frequency effect for new test probes in the CM condition: the predictions for the HF items lie virtually on top of the predictions for the LF items for both the RT and accuracy data. By comparison, in the observed CM data, the RTs for the HF new items are much shorter than for the LF new items. Because the familiarity boost from HF items should be greater than for LF items, if anything the FAM model would predict that RTs for HF new items should be *longer* than for LF items, not shorter. (Its prediction of equality is achieved only by setting the FAM_{HF} and FAM_{LF} parameters equal to one another – see Table 3.)

In addition, the FAM model struggles to account for the data from the VM condition. Although it correctly predicts the HF disadvantage for the new test probes, it failed to predict the trend of an HF disadvantage for the old test probes in the observed data (for both the RTs and the error probabilities). According to the model, when an HF item serves as a test probe, it will receive a higher familiarity boost from LTM (compared to LF items). If anything, this boost should *facilitate* responding "old"; thus, the model predicts somewhat shorter RTs and increased

accuracy for HF old items than LF old items. By contrast, the observed data tend to show higher error rates and slightly longer RTs for the HF old items than for the LF old items. Because the FAM model failed to account for the data even with all its free parameters allowed to vary across conditions, we did not explore more constrained versions of the model.

In contrast to the FAM model, the full version of the IR model successfully captured most of the data patterns (and provided a better fit than the FAM model to both the CM and VM data -- see Table 2). However, because a large number of free parameters were used, we decided to explore a series of more constrained versions of the IR model that might still achieve good accounts of the data. In each case, we held fixed across the CM and VM conditions additional parameters (rather than allowing the parameters to vary freely across the conditions). As can be seen in Table 2, with each additional constraint, there was a relatively small increase in the total WSSD. Here we focus on the most constrained version (the "core" IR model). In this version, we held fixed across the CM and VM conditions: the scale κ and residual time T_r parameters; the similarity parameter *s*; the lag-decay parameter β; and the criterion parameters *u* and *v*. In addition, although one might expect the response-boundary parameters to vary in magnitude across both conditions and response types, we found that reasonable fits could be achieved with all response-boundary boundary parameters set at a single value. Thus, the key parameters that vary across conditions are the various LTM parameters (*IR-OLD*$_{HF}$, *IR-OLD*$_{LF}$, *IR-NEW*$_{HF}$, and *IR-NEW*$_{LF}$) -- see Table 3.

The predictions of the mean RTs and error probabilities from the core IR model are presented in Figures 1C and 2C. (In addition, we show the predictions for the more fine-grained set-size by lag curves in Figures 3B and 4B.) Although there are some minor exceptions, the model successfully captures most of the main qualitative effects in the data.[3] In this discussion,

we focus on the effects of the frequency manipulation, the central theme of the present investigation. To begin, the model captures the overall HF advantage in the CM condition – for both old *and* new probes. According to the model, the test probe will receive a boost toward the correct response boundary when it activates the "item-plus-response-label" exemplars from past trials, because the response label is consistent with the correct response on the current trial. The magnitude of the boost should be positively correlated with the frequency of the consistent mapping (see also Schneider & Fisk, 1982, who manipulated *proportion* of CM trials associated with individual items in visual-search paradigms). This boost leads to the better performance for HF items than for LF items for both "old" and "new" test probes.

More surprising is that, compared to the FAM model, the IR model also provided a better account of the VM data. In particular, the IR model successfully predicted that performance on the old *HF* test probes would tend to be worse than on the old *LF* test probes. According to the model, although an old HF test probe will receive a strong boost toward +*OLD* from past trials, it will also receive a strong boost toward –*NEW* (from the many trials in which it served as a new test probe). As we explained previously, this strong LTM-based interference from past trials adds noise to the random-walk process (as formalized in the Equation-8 step-probability equation). There is less noise from LTM added for the LF items, so performance is governed more by the contents of the current study list.

The best-fitting parameters from the core IR model are reported in Table 3 and the pattern of parameter estimates seems reasonably interpretable. Naturally, for each response type (OLD vs. NEW), the LTM parameter values (IR-OLD and IR-NEW) associated with the HF items are greater in magnitude than those associated with the LF items: we imposed this relation as a constraint in our model fitting, but it emerged clearly in any case. In addition, for CM, the

LTM parameters associated with old items were greater in magnitude than the LTM parameters associated with new items. A natural explanation is that, in addition to their consistent response mappings throughout the experiment, the old items had the advantage of often appearing on the study lists, whereas the new items would never appear on the study lists. The very low-magnitude estimate for IR-NEW_{LF} also seems reasonable, because individual LF items rarely appeared as new test probes, so there was little opportunity for item-response learning to occur for these items.

Another general result of the parameter estimates for the IR model is that the magnitude of the LTM parameter values for CM tended to be greater than the magnitude of the LTM parameters for VM. There are several reasonable explanations. Perhaps most important, in a fully specified model, provision would be made for the observer to differentially weight the STM and LTM sources of information in making old-new recognition judgments. An effective observer would give far more weight to LTM in CM compared to VM. The reason is that the LTM item-response mappings are perfectly valid for the CM paradigm; whereas they provide zero information in the VM paradigm. Indeed, in VM, an effective observer would attempt to ignore the previous-lists history and focus solely on the contents of the current list, perhaps through the use of recency-based context cuing (e.g., Howard & Kahana, 2002). Such context cuing would tend to zero out the influence of memory traces laid down in the distant past. In sum, the differing magnitudes of the LTM parameters across CM and VM may be reflecting the greater weight that observers give to LTM in the CM condition than in the VM condition.[4]

A second issue involving the differing magnitudes of the LTM parameters across CM and VM is that the detailed mechanisms for the development of the item-response strengths remain to be delineated. For example, although one possibility is that observers simply accumulate

individual item-response exemplars as they are experienced, an alternative is that in VM the "old" and "new" trials compete with one another to some extent, weakening the learned item-response strengths of both. Third, although the absolute frequency of individual item types was roughly balanced across the CM and VM conditions, it was of course the case that the frequency of assignment of items to specific responses differed: In CM an HF item was either always assigned to an old response or always assigned to a new response, whereas in VM an HF item was assigned roughly half the time to each response. Future research will be needed to disentangle these alternative explanations of the differing magnitude of the CM and VM long-term-memory parameters.

Finally, to gain additional information bearing on the nature of the LTM mechanisms in these tasks, we created summary plots of the data patterns shown in Figures 1 and 2, separately for Blocks 2-3 and Blocks 4-5 of the experiment. Visual inspection indicated very similar patterns of results across these early and late blocks, suggesting that the detailed mechanisms that give rise to the LTM parameter values across CM and VM have their influence at fairly early stages of practice.

Discussion

Our results suggest that item-response learning is a core mechanism through which long-term memory (LTM) influences performance in tasks of both CM and VM short-term probe-recognition. According to this view, each presentation of a test probe leads to the storage of that test probe – along with its associated "old" or "new" response -- as an exemplar in LTM. These

item-response pairs are retrieved along with current-list items in driving observers' short-term probe-recognition judgments.

The key manipulation used to diagnose the influence of this LTM mechanism in the present experiment was to vary long-term item frequencies across trials. In CM, increased item frequency led to better performance for both old and new items. Although the improved performance on old items is predicted by both item-response-learning and item-familiarity mechanisms, the improvement on the new items suggests a crucial role of item-response learning. In particular, item-familiarity mechanisms predict that increased frequency should have interfered with responding "new" to new items, but the results pointed decidedly in the opposite direction. Instead, our theoretical interpretation is that retrieval of the high-frequency new-item-response pairs from LTM boosted the evidence that such test probes were new. These results provide converging evidence for the significant role of item-response learning in CM memory search reported in the earlier study of Nosofsky, Cao et al. (2014), in which the key manipulation was to repeat test probes across successive trials. However, whereas the evidence from that previous study may have reflected paradigm-specific strategies in which observers became sensitized toward the repeating-probe manipulation, the present results provide evidence of a much more general long-term item-response-learning influence on CM memory-search performance.

Whereas increased item frequency led to facilitation in CM memory search, it led to interference (for both new and old items) in VM memory search. Although the interference for new test probes is as predicted by familiarity-only models, the interference for old items contradicts the predictions from such models. In particular, because increasing frequency boosts familiarity, a familiarity-only model predicts that there should have been benefits in responding

"old" for the high-frequency items, not interference. Instead, the overall pattern of results, even for VM, was again well explained by an item-response learning model. The basic idea is that in the VM condition there were numerous previous trials in which the high-frequency items were assigned both old and new responses. Retrieval of these numerous conflicting item-response pairs adds significant noise to the decision process, regardless of whether the correct response on the current trial is old or new. A formal implementation of this mechanism within an exemplar-based random-walk model yielded good qualitative accounts of the overall pattern of results across both the CM and VM conditions.

Thus, the present empirical and formal-modeling results suggest the possibility of considerable generality for the role of an item-response-learning component in wide varieties of memory-search performance. We suspect that a complete model of how memory-search performance evolves with practice will involve the contribution of multiple mechanisms, and we certainly do not conclude that long-term item-familiarity mechanisms do not also play a role. In addition, future work will also need to examine the extent to which the performance patterns for the high-frequency items reflect explicit strategies developed for those items or reflect processes that are more hard-wired into the memory system. Regardless of the outcome of these future investigations, the present evidence suggests that, at the least, item-response-learning mechanisms will form a core component of a fully satisfactory unified model of CM and VM memory search.

References

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2,* 396-408.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, USA, 105,* 14325– 14329.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436.

Donkin, C., & Nosofsky, R. M. (2012a). A power-law model of psychological memory strength in short-and long-term recognition. *Psychological Science, 23,* 625-634.

Donkin, C., & Nosofsky, R. M. (2012b). The structure of short-term memory scanning: An investigation using response time distribution models. *Psychonomic Bulletin & Review, 19,* 363-394.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46,* 269-299.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95,* 492–527.

Logan, G. D. (1990). Repetition priming and automaticity: Common underlying mechanisms?. *Cognitive Psychology, 22,* 1-35.

McElree, B., & Dosher, B. A. (1989). Serial position and set size in short-term memory: Time course of recognition. *Journal of Experimental Psychology: General, 18,* 346–373.

Monsell, S. (1978). Recency, immediate recognition memory, and reaction time. *Cognitive Psychology, 10,* 465–501.

Nosofsky, R. M. (2016). An exemplar-retrieval model of short-term memory search: Linking

categorization and probe recognition. *Psychology of Learning and Motivation, 65*, 47-84.

Nosofsky, R. M., Cao, R., Cox, G. E., & Shiffrin R. M. (2014). Familiarity and categorization processes in memory search. *Cognitive Psychology, 75*, 97-129.

Nosofsky, R. M., Cox, G. E., Cao, R., & Shiffrin, R. M. (2014). An exemplar-familiarity model predicts short-term and long-term probe recognition across diverse forms of memory search. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 1524.

Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review, 188*, 280–315

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*, 266–300

Nosofsky, R. M., & Stanton, R. D. (2005). Speeded classification in a probabilistic category structure: Contrasting exemplar-retrieval, decision-boundary, and prototype models. *Journal of Experimental Psychology: Human Perception and Performance, 31*, 608-629.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108.

Schneider, W., & Fisk, A. D. (1982). Degree of consistent training: Improvements in search performance and automatic process development. *Perception & Psychophysics, 31*, 160–168.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84*, 1–66

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84,* 127–190.

Strayer, D. L., & Kramer, A. F. (1990). An analysis of memory-based theories of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 291–304.

Strayer, D. L., & Kramer, A. F. (1994). Strategies and automaticity: I. Basic findings and conceptual framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 318–341.

Sternberg, S. (1966). High-speed scanning in human memory. *Science, 153*, 652– 654.

Sternberg, S. (2016). In defence of high-speed memory scanning. *The Quarterly Journal of Experimental Psychology*, *69*, 2020-2075.

Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science, 2*, 409-415.

Wolfe, J. M., Boettcher, S. E., Josephs, E. L., Cunningham, C. A., & Drew, T. (2015). You look familiar, but I don't care: Lure rejection in hybrid visual and memory search is not based on familiarity. *Journal of Experimental Psychology: Human Perception and Performance, 41*, 1576.

Footnotes

1. Although there are some exceptions, it is generally the case that the frequency effects are larger at the large memory set sizes than at the smaller ones. This pattern of results is not surprising because there will be tend to be floor effects for small-size memory sets due to the ease of those conditions. Because our main concern is with overall item-frequency effects across the CM and VM conditions, in order to facilitate reading of our Results section, we report the outcome of the more detailed tests of set-size effects and tests of interactions between set size and item frequency in table form in the appendix.

2. As noted in our introduction, different strategies may operate in the short-term probe-recognition task depending on details of the experimental conditions (Sternberg, 2016). The strategy formalized in the EBRW model is hypothesized to operate under conditions involving fairly rapid presentations of memory-set items, short intervals between study and test, and no requirement that the participants report the ordering of the memory-set items subsequent to their old-new judgment on each trial.

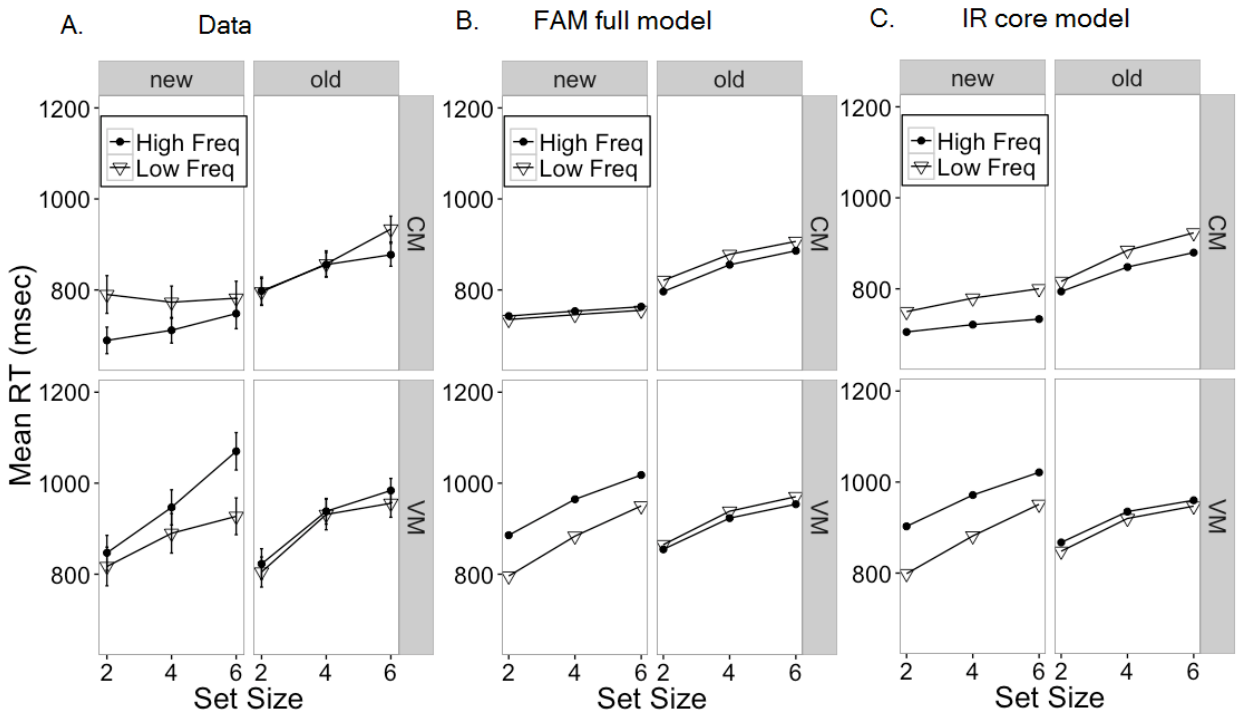
3. One minor exception is that the current version does not predict the decreases in RTs and error rates that are often observed at the greatest lag of each set-size condition. This decrease constitutes a primacy effect: the item with the greatest lag occupies the first serial position of the memory set. In past applications of the EBRW-recognition model (e.g., Nosofsky et al., 2011), a primacy parameter was added to capture the effect, but the effect seems tangential to the main issues under investigation in the present study. A second possible limitation is that the present

version of the model predicts little if any additional effect of memory set size once once one conditions on lag. In past applications, any small residual effects of set size on RT have been modeled in terms of increases in response-threshold settings (e.g. Donkin & Nosofsky, 2012b; Nosofsky et al., 2011; see also Ratcliff, 1978). Again, however, this more detailed issue is not central to the present investigation.

4. We should note that this differential-weighting hypothesis is reminiscent of the type of dual-process theory developed by Shiffrin and Schneider (1977) to account for the contrasting results across the CM and VM conditions of their hybrid visual/memory-search studies. For example, consider a simple form of dual-process model: CM, after enough training, would be accomplished by memorized and/or automatic stimulus-response pairings without retrieval or use of recently stored study or test traces. VM would be carried out with the use of context cuing that would activate recent traces, with responses based on some combination of item information and item-response information retrieved from those traces. (Of course, such a simple model for CM in our paradigm is unlikely because low-frequency items do not receive much training, and subjects likely cannot help attending to the study lists, so that even from a dual-route perspective CM performance is likely a mixture of both routes.) Our present studies were not designed to distinguish between the type of single-process model presented in this article and the type of dual-process theory described by Shiffrin and Schneider (1977) and future research would be needed to distinguish between such possibilities.

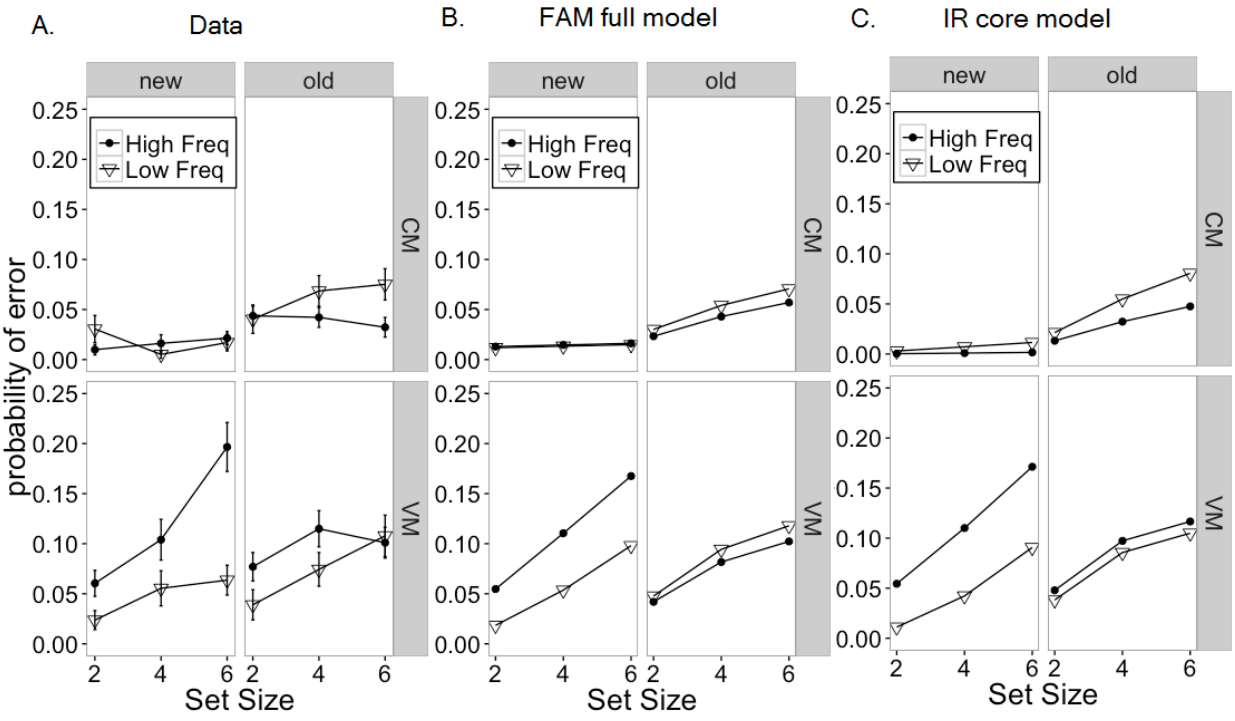
Figures

Figure 1



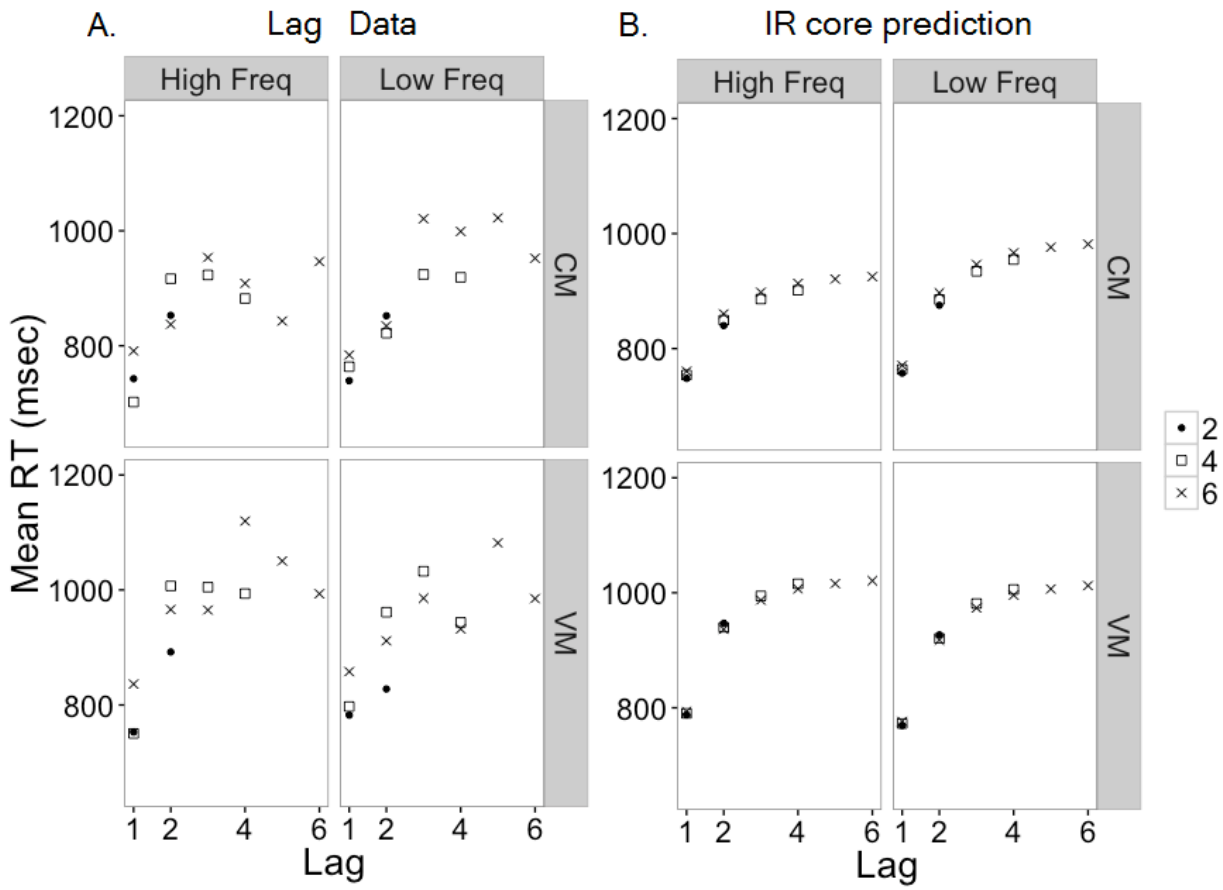
Mean correct response times plotted as a function of conditions (CM, VM), set size, test-probe type (new, old), and item frequency (HF, LF). Left panel = observed data, middle panel = predictions from full version of item-familiarity model, right panel = predictions from core version of item-response-learning model. Error bars show the between-subject standard-error of the mean.

Figure 2



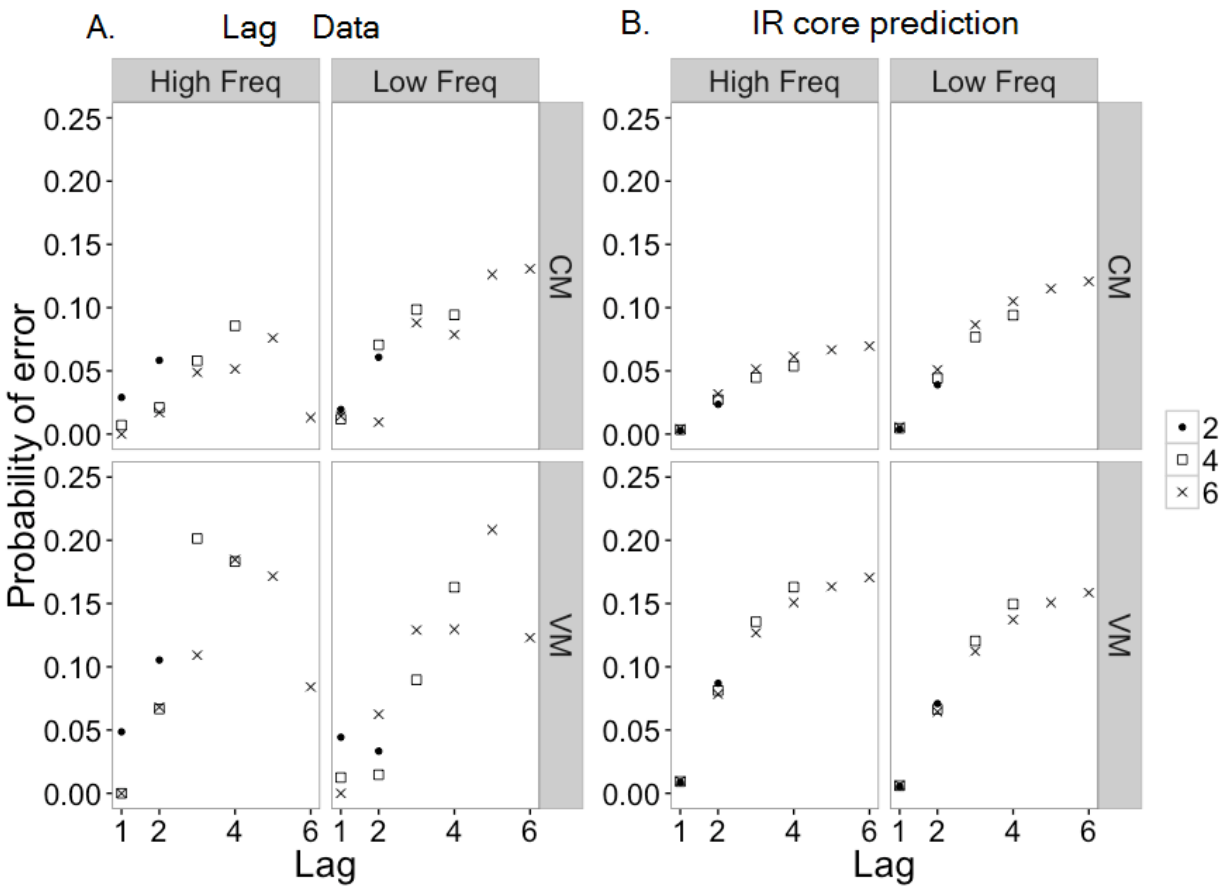
Mean error probabilities plotted as a function of conditions (CM, VM), set size, test-probe type (new, old), and item frequency (HF, LF). Left panel = observed data, middle panel = predictions from full version of item-familiarity model, right panel = predictions from core version of item-response-learning model. Error bars show the between-subject standard-error of the mean.

Figure 3



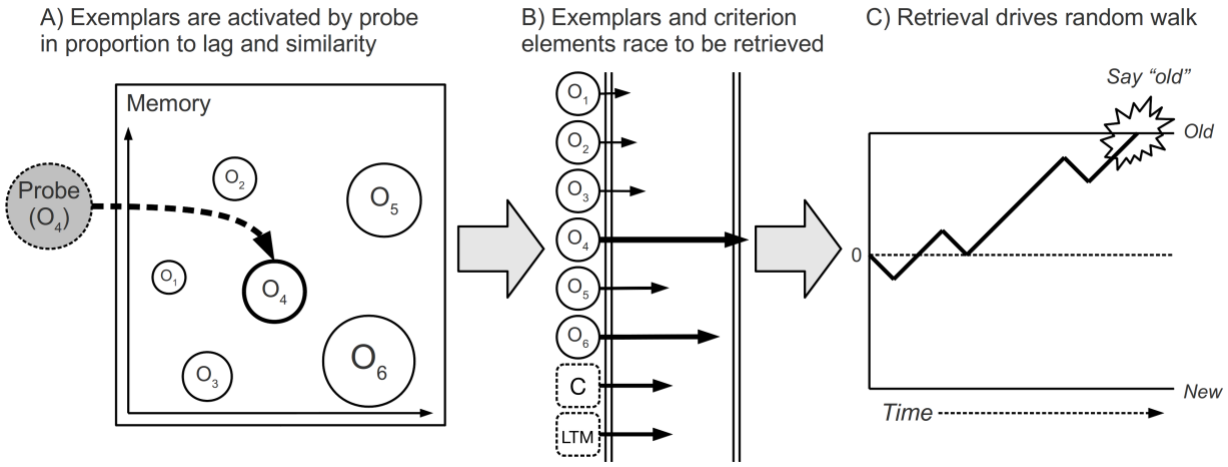
Mean correct response times for the old test probes plotted as a function of conditions (CM, VM), set size, lag, and item frequency. Left panel = observed data, right panel = predictions from core version of item-response-learning model.

Figure 4



Mean error probabilities for the old test probes plotted as a function of conditions (CM, VM), set size, lag, and item frequency. Left panel = observed data, right panel = predictions from core version of item-response-learning model.

Figure 5



Schematic illustration of the application of the exemplar-based random-walk model to the short-term probe-recognition task. Note: O_k is the old item on the current study list that is presented in serial-position k.

Tables

Table 1. Relative Frequency for individual HF, MF and LF items

	Prob. OLD test probe	Prob. NEW test probe	Prob. Tested	Prob. Studied
CM-positive HF	0.095	0	0.095	0.708
CM-positive MF	0.060	0	0.060	0.496
CM-positive LF	0.016	0	0.016	0.133
CM-negative HF	0	0.118	0.118	0
CM-negative MF	0	0.060	0.060	0
CM-negative LF	0	0.012	0.012	0
VM HF	0.053	0.044	0.097	0.43
VM MF	0.031	0.029	0.060	0.25
VM LF	0.007	0.008	0.017	0.05

CM=Consistent Mapping; VM=Varied Mapping; HF=High Frequency; MF=Medium Frequency; LF=Low Frequency. Prob. = Probability.

Table 2 Weighted Sum of Squared Deviation (WSSD) Fits of Different Versions of the FAM and IR Models to the Mean Correct RTs and Error-Probability Data

Model	**VM**	**CM**	**Total**
FAM full model	0.181	0.113	0.294
IR full model	0.159	0.071	0.230
IR κ, T_r	0.160	0.072	0.224
IR κ, T_r, *s*	0.161	0.075	0.236
IR κ, T_r, *s*, *u*, *v*	0.164	0.076	0.240
IR κ, T_r, *s*, *u*, *v*, +OLD, -NEW	0.165	0.081	0.246
IR core	0.163	0.085	0.248

FAM = item-familiarity model, IR = item-response learning model, VM = varied-mapping, CM = consistent-mapping. The parameter listings next to the IR model denote the parameters that were held fixed across the VM and CM conditions in fitting the special-case versions of the model to the data. (See Table 3 for a listing and description of the parameters.) The IR-core model was the most highly constrained of the special-case IR models and held fixed across VM and CM the parameters κ, T_r, *s*, *u*, *v*, +OLD, -NEW, and β.

*Table 3.*The best-fitting parameter values for full FAM model and core IR model

	FAM full			IR core		
Parameters		**CM**	**VM**		**CM**	**VM**
α		5.420	0.468		0.174	0.456
β		0.934	1.870		2.057	-
u		27.813	0.324		0.394	-
v		0.334	0.010		0.015	-
s		0.059	0.043		0.065	-
	*FAM_{HF}	22.816	0.043	*$IR\text{-}OLD_{HF}$	0.557	0.115
LTM				$IR\text{-}OLD_{LF}$	0.433	0.002
Parameters	FAM_{LF}	22.7254	0.001	*$IR\text{-}NEW_{HF}$	0.298	0.088
				$IR\text{-}NEW_{LF}$	0.001	0.001
OLD		22.854	3.125		4.109	--
NEW		107.325	2.874		--	--
κ		0.349	84.907		36.037	-
Tr		349.695	378.527		507.697	-

Note. FAM = item-familiarity model, IR = item-response learning model, CM = consistent-mapping, VM = varied-mapping. Parameter values replaced with " – " were set equal to one another across the CM and VM conditions; all response thresholds were held fixed at a single value in the IR core model. Parameters marked with "*" were constrained to be greater than or equal to the parameter immediately below. α = power-decay asymptote, β = power-decay rate, *u* = criterion intercept, *v* = criterion slope, *s* = similarity, OLD = old threshold, NEW = new threshold, κ = scale, Tr = residual time. See text for an explanation of the LTM parameters.

Appendix

Table of statistical-test results examining main effects of set size and interactions between set size and item frequency for both RT and error probability for CM-old, CM-new, VM-old, and VM-new.

	Set Size	**Frequency**	**Interaction**
CM old	**F (2,102) = 13.76**	**F (1, 51) = 4.59**	F (2, 102) = 0.22
RT	**P < 0.001**	**P = 0.037**	P = .801
CM old	F (2, 102)=0.29	**F (1, 51) = 8.26**	F (2, 102) =1.03
p (error)	P = 0.750*	**P = 0.006**	P=0.359*
CM new	F (2, 102)=1.30	**F (1, 51) = 12.96**	F (2, 102) = 2.69
RT	P = 0.276	**P < 0.001**	P = 0.073
CM new	F (2, 102)=1.02	F (1, 51) = 0.07	F (2, 102) = 2.43
p (error)	P=0.364	P = 0.793	P = 0.093
VM old	**F (2, 100)=40.70**	F (1,50) = 0.02	F (2, 100) = 0.17
RT	**P < .001**	P = 0.884	P = 0.842
VM old	**F (2, 100)=6.31**	F (1,50) = 1.84	**F (2, 100) = 3.86**
p (error)	**P = .003**	P = 0.182	**P = 0.024**
VM new	**F (2, 100)=32.71**	**F (1, 50) = 18.43**	F (2, 100) = 4.84
RT	**P < 0.001**	**P < 0.001**	P = 0.010*
VM new	**F (2, 100)=24.48**	**F (1, 50) = 29.69**	**F (2, 100) = 7.14**
p (error)	**P < 0.001***	**P < 0.001**	**P = 0.001**

Note. Asterisks denote cases in which there were violations of sphericity. Boldface entries denote cases that are statistically significant at least the p = .05 level. All analyses conducted in JASP

Is Item-Response Learning Strategy Independent?

Probe-recognition memory search tasks are one of the most important paradigms for understanding memory. In these tasks, on each trial, subjects are presented with a list of to-be-remembered items (the "memory set"). The memory set is followed by a test probe that is either a member of the memory set (an *old* item or *target*) or not (a *new* item or *foil*). Subjects are asked to judge whether the test probe is a target or a foil as rapidly as possible without making errors. It is generally observed that response times (RTs) get longer and accuracy decreases as the size of the memory set grows, a pattern termed the set-size effect. The set-size effect was first reported in Sternberg (1966), and replicated in many follow-up studies that investigated the underlying processes of memory search tasks (e.g., McElree & Dosher, 1989; Monsell, 1978; Nosofsky, et al., 2011). Although the detailed assumptions of the proposed processes might differ depending on the specific experimental conditions (for a comprehensive review and analysis, see Sternberg, 2016), all the theories assumed that subjects engage the current list items in short-term memory to perform the task. Therefore, longer lists result in worse performance since short-term memory is generally assumed to be capacity-limited (Atkinson & Shiffrin, 1968).

In their studies that examined hybrid forms of memory/visual search, Schneider and Shiffrin (1977) discovered that the set-size effect could be greatly reduced or eliminated under "consistent mapping" (CM) conditions. In CM, the target probes are chosen from a fixed set on every trial, and the foil probes are chosen from a different fixed set on every trial. Thus, the targets and the foils never switch roles across trials. As practice proceeded in Schneider and Shiffrin's studies, performance improved dramatically: subjects were able to make their old/new

judgments with decreasing RT and few errors. Most importantly, the performance became largely invariant to the set-size manipulation, suggesting reliance on a process other than the retrieval of the list held in short-term memory (see also Logan & Stadler, 1991). In Schneider and Shiffrin (1977), subjects were also tested in a varied-mapping (VM) condition, where the items that served as old probes on some trials were new probes on other trials, and vice versa. In contrast to CM, performance in the VM condition improved very little with practice, and the set-size effect persisted even after extensive practice. The researchers proposed that performance in the VM condition required an effortful, controlled process, regardless of the amount of practice; whereas practice in the CM condition allowed for the development of an extremely efficient, automatic form of information processing.

Despite the numerous studies aiming to explore the nature of automaticity (i.e. Cao, Nosofsky, Shiffrin, 2016; Cheng, 1985; LaBerge & Samuels, 1974; Logan & Stadler, 1991; Schneider & Fisk, 1982; Shiffrin and Schneider, 1977; for review, see Schneider & Chein, 2003), the precise cognitive mechanisms that underlie the development of automaticity remain unclear. Specifically, the manner in which skilled performance develops in CM training and whether the learning is strategy dependent are issues that are not well understood. To address such questions requires the development of a quantitative model that explains *not only* how CM training improves performance *but also* how VM training impacts performance. One influential model that aims to provide a quantitative account of the development of automaticity is Logan's (1988) instance theory. According to instance theory, highly efficient automatic performance arises by retrieving responses that are linked to the instances stored in long-term memory. With increased practice, more instances are stored in memory, which leads to a more efficient retrieval process (see Logan, 1988, 1990 for details). Furthermore, Logan (1988) suggested that the

dramatic performance improvement observed in CM conditions was an automatic byproduct of the memory system.

Some progress towards a quantitative model for both CM and VM performance was made in recent work by Nosofsky and colleagues (Cao, Shiffrin, & Nosofsky, 2018; Nosofsky, Cao, Cox, & Shiffrin, 2014; Nosofsky, Cox, Cao, & Shiffrin, 2014). The formal model is an extended version of the exemplar-based random-walk (EBRW) model that has been successfully applied to various forms of categorization (Nosofsky & Palmeri, 1997; Nosofsky & Stanton, 2005) and old–new recognition memory (Nosofsky et al., 2011). In the version of the model applied to probe recognition memory search, each item of the memory set is stored as an exemplar in short-term memory. Items from previous memory sets may also be stored in long-term memory. When the test probe is presented, it activates exemplars to which it is similar (both short-term and long-term), and the activated exemplars race to be retrieved (see Formal Models section for details). The retrieved exemplars are evaluated and lead the evidence accumulator to move toward either the "Old" response threshold or the "New" response threshold. The retrieval process continues until one of the thresholds is reached and the observer then emits the associated response.

Nosofsky and colleagues identified two different processes that the retrieved exemplars could contribute as evidence toward the old/new judgment. In an "item-familiarity" (termed FAM) version of the model, a retrieved exemplar (from short-term or long-term memory) leads the information accumulator toward the "old" threshold, while failure to retrieve an old exemplar leads toward the "new" threshold. By contrast, the "item-response" (termed IR) version of the model assumes that subjects learn and store the response label associated with each test probe along with that test probe, and later rely on the stored response label to perform the task. After

sufficient learning, retrieval of exemplars with "old" response labels would drive the random walk toward the old response threshold, but retrieval of exemplars with "new" response labels would drive the information accumulator toward the "new" response threshold. The results produced strong evidence that item-response learning plays a key role in CM conditions-- the IR model provided excellent quantitative accounts of CM performance in various experiments (for details, see Nosofsky, Cao, et al. 2014 and Cao, et al. 2018).

Moreover, Nosofsky and colleagues found evidence suggesting that item-response learning may impact performance in the VM conditions, a surprising result to many researchers and theorists who thought the inconsistent training in VM would prevent learning and possibly prevent storage of IR responses in long-term memory. In an experiment manipulating the frequency of individual items in VM conditions, increased item frequency did *not* facilitate performance when subjects should respond "old" to the test probe (Cao, et al. 2018). Such a result challenges the familiarity-only model because the model predicts a benefit in responding "old" for the high-frequency items. On the other hand, this result could be explained by item-response theory using an assumption that the numerous previous trials in which the high-frequency items were assigned to both old and new responses add significant noise to the decision process, regardless of whether subjects should respond "Old" or "New" on the current trial. The fact that the IR model provided an excellent account for VM performance, even though keeping track of response labels hinders performances in VM, suggests that IR learning could be a default learning mechanism underlying both CM and VM training.

The present study further explores the possibility that item-response learning process is the core mechanism underlying VM and CM memory search. The exploration is based on new empirical data and computational modeling. The experiments discussed earlier varied CM and

VM in different blocks of training or on different trials, raising the possibility that different strategies could have been at work. Strayer and Kramer (1994a) reported reduced differences between CM and VM when the two types of items were mixed within the same trials. They characterized the performance through use of diffusion modeling (Ratcliff, 1978) but did not aim to provide a mechanistic account of the cognitive processes that lead to the differences.

I therefore conducted two experiments that mixed VM and CM items within trials, and applied IR and FAM models to the findings. I had three primary questions: First, would mixing reduce the large differences we observed for CM and VM training. The simplest form of Logan's (1988) theory predicted CM learning would occur regardless of tasks and conditions and therefore would cause large advantages for CM even in mixed conditions. The results of Strayer and Kramer (1994a) seemed at odds with this prediction but I wanted to obtain additional evidence within our paradigm. Second, although Strayer and Kramer (1994a) could characterize their results with diffusion modeling, it is unclear if a process-based model, particularly the IR model, would be able to account for the findings with strategic parameters fixed for both item types. Third, could the mixed conditions produce a switch of process away from the use of IR and to the use of familiarity? Because item-response learning is a suboptimal strategy for VM items, mixing might cause a reversion to use of the FAM model.

In the remainder of this article, I first review the formal EBRW model and explain how the IR and FAM mechanisms are implemented in the different versions of the model. Next, I report two experiments where different types of items were mixed within the same memory sets on each trial. In the first experiment, CM items were mixed with VM items. In the second experiment, I mixed CM and VM items with items that were never presented in previous trials ("all-new" items) to further test the models. As shown in the model-fitting results, the study

revealed a much more complicated story that challenges the simple view that CM and VM performance arises as an automatic consequence of IR learning. Instead, CM and VM performance appears to also be highly dependent on task-specific strategic factors.

The Formal Model

A schematic illustration of some of the main components of the EBRW is presented in Figure 1. First, consider the study items from the memory set of the current trial. According to the model, each of the study items is stored in memory as an individual exemplar. The memory strength of the exemplar is presumed to decay solely as a function of the lag with which it was presented on the list (with the most recently presented items having the shortest lags). Based on the evidence reported by Donkin and Nosofsky (2012), we further assume that the memory strength decreases as a power function of lag *j*:

$$m_j = \alpha + j^{-\beta}, \qquad (1)$$

where β reflects the rate of decrease and α reflects asymptotic strength at long lags. Because larger-size memory sets have items with long lags, this component of the model helps explain why probe-recognition performance tends to get worse as memory set-size increases (see below). The differential memory strengths are represented schematically in Figure 1 (panel A) where the larger circles represent exemplars with greater memory strength.

When the test probe is presented, exemplars stored in memory are "activated" and "race" to be retrieved, with rates that are proportional to their activations (cf. Logan, 1988), as illustrated in panel B of Figure 1. The degree to which exemplar j (e_j) is ''activated'' by test-item i (t_i) is a joint function of exemplar j's memory strength and its similarity (given by a free parameter s, $0 < s < 1$) to test item i:

$$a_{ij} = m_j, \quad \text{if } t_i = e_j \tag{2a}$$

$$a_{ij} = m_j s, \quad \text{if } t_i \neq e_j \tag{2b}$$

Thus, the study-list exemplars that are most highly activated are those that match the test probe and that have short lags.

To apply EBRW to recognition tasks, a process that provides evidence for "new" items needs to be implemented. In most cases, such a process is instantiated in recognition models by assuming that subjects set a "criterion": if the evidence for "Old" response exceeds the criterion, the information-accumulation process moves toward the "Old" response threshold; whereas if the evidence fails to exceed the criterion, the information-accumulation process moves in the direction of a "New" response (Ratcliff, 1985). Here, to adapt the evidence-criterion setting process described above to the exemplar-retrieval model, we make an analogous assumption. In particular, we assume that the observer establishes what we term "criterion elements" in the memory system. Just as in case for the stored exemplars, upon presentation of a test probe the criterion elements (labeled "c" in Figure 1B) race to be retrieved. Opposite to the old exemplars, retrieval of the criterion elements leads the random-walk process to step toward the "New" response threshold. Just like the evidence-criterion setting process, a higher activation value for criterion elements results in a lower probability of old exemplars winning the race, so the random-walk process is more likely to step toward the "New" response threshold.

Finally, the retrieved exemplars and criterion elements drive a random-walk process that determines the "Old" vs. "New" decisions (see Figure 1 panel C). The observer sets response thresholds +*OLD* and -*NEW* that establish the amount of evidence needed for making an "Old" or a "New" response. On each step of the random-walk process, if an old exemplar is

successfully retrieved, the random-walk counter takes a step toward the "Old" response threshold; whereas if a criterion element wins the race, the random-walk counter take a step toward the "New" response threshold. The retrieval process continues until one of the response thresholds is reached, at which point the observer carries out the appropriate response.

Given the assumptions described above (and some further technical assumptions described by Nosofsky and Palmeri, 1997), it turns out that, on each step of the random-walk, the probability that the random-walk counter steps toward the +*OLD* response threshold is given by:

$$p_i = A_i / (A_i + k), \qquad (3)$$

where A_i represent the summed activation of the test probe to all the study list items:

$$A_i = \sum a_{ij}, \qquad (4)$$

and k is the level of criterion-element activation. (The mean rate of accumulation to the -*NEW* response boundary is given by $q_i = 1 - p_i$.)

Through experience in the task, the observer is presumed to learn an appropriate setting of the criterion-element activation k, such that the summed activation (A_i) tends to exceed k when the test probe is old, but tends to be less than k when the test probe is new. Because A_i tends to increase with set size for new test probes, allowance is made for the possibility that the observer adjusts k with increases in set size. As an approximation, it is assumed that the criterion setting varies linearly with memory set-size M:

$$k(M) = u + v \cdot M. \qquad (5)$$

The development thus far has considered the role of only the items on the current study list. However, a key to providing a full explanation of VM vs. CM performance is to also formalize the role of the study and test items presented on previous trials of the experiment.

Our current model implements the influence of past trials with a set of long-term memory (LTM) components (see Figure 1B). In theory, when the test probe is presented, it will activate all the traces stored in long-term memory to various degrees. For simplicity, however, we assume that only LTM exemplars that match the test probe may be retrieved and lead the random-walk counter to step toward a response threshold. According to one version of the model, when a LTM exemplar is retrieved, it influences performance in the same way as retrieved exemplars from the current study list, so the random-walk counter moves toward the *+OLD* threshold, regardless if the LTM exemplar was originally studied as "Old" or tested as "New" test probe. We refer to this version of the model as an *item-familiarity* (*FAM*) model and formalize the model with a set of *FAM* parameters. Alternatively, according to a second version of the model, the observer stores the response labels associated with the test probes from previous trials. If a LTM exemplar with an "Old" response label is retrieved, it will lead the random-walk counter to step toward *+OLD;* whereas if an LTM exemplar that served as a "New" test probe is retrieved, it will lead the random-walk counter to step toward *-NEW.* We refer to this second version of the model as an *item-response-learning* (*IR*) model and formalize it with a set of *IR* parameters. The details of both models are described below.

LTM-FAM

In the FAM model, we presume that the activation and retrieval of LTM exemplars always leads the random-walk to step toward *+OLD.* For simplification, we account for the boost in the summed activation (A_i) with a free parameter *FAM:*

$$p_i = (A_i + FAM) / [(A_i + FAM) + k]. \quad (6)$$

As will be seen in the model-fitting sections, the magnitude of FAM may depend on the type of item that is tested.

LTM-IR

In the IR model, we presume that the retrieved LTM exemplar-plus-response-label will direct the random-walk counter to the corresponding response threshold. The boost in activation toward +*OLD* that is produced by retrieved exemplars with old response-labels is denoted *IR-OLD;* whereas the boost in activation that is produced by retrieved exemplars with new response labels is denoted *IR-NEW.*

In CM conditions, a target will activate many exemplars with "Old" response-labels and no exemplars with "New" response-labels. Thus, for CM targets, the probability that the random walk steps toward +*OLD* is given by:

$$p_i \text{ (old)} = (A_i + IR\text{-}OLD) / [(A_i + IR\text{-}OLD) + k]. \quad (7a)$$

Conversely, a test foil will activate previous test probes with "New" response labels but none with "Old" response labels. Thus, for CM foils, the probability that the random walk steps toward –*NEW* is given by

$$q_i \text{ (new)} = (k + IR\text{-}NEW) / [(k + IR\text{-}NEW) + A_i]. \quad (7b)$$

In VM conditions, a test probe will activate exemplars with "Old" response-labels as well as exemplars with "New" response-labels from past trials, regardless if it serves as an old or new

test probe on the current trial. Therefore, regardless of item type, the probability towards *+OLD* is given by:

$$p_i = (A_i + IR\text{-}OLD) / [(A_i + IR\text{-}OLD) + (k + IR\text{-}NEW)], \quad (7a)$$

As is the case for the FAM model, in applying the IR model to the experiments, the magnitude of the IR-OLD and IR-NEW parameters will be allowed to depend on the specific item types that are tested (see Model-Fitting sections for details).

Note that in both the IR and FAM models, one cannot assume that strategy-related parameters influence only the settings of the response thresholds, as proposed in Strayer and Kramer (1994a). For example, an observer could adjust the criterion-element parameters in response to the different set-size manipulations, which changes the probability of different directions of the random-walk (drift rate), while not reflecting any long-term changes in memory representations. Likewise, the use of FAM vs IR processes may be strategy-dependent and could vary with conditions of testing. Both the FAM and IR models formalize more detailed mechanisms of how changes in strategy and changes in memory representations may influence CM and VM performance than was provided by the simplified assumptions in Strayer and Kramer (1994a).

Experiment 1

We tested subjects in a probe-recognition memory-search task, with both CM and VM training. In the task, we mixed CM and VM items within trials. We also manipulated the size of the memory sets across trials. The behavior pattern should provide evidence to help answer the question of whether the item-response learning mechanism plays a major role in the mixed condition. Particularly, under the FAM model assumption, there should be little difference in

performance between the VM targets and CM targets because both were presented frequently during training. By contrast, the IR model would predict substantially worse performance for the VM targets than for the CM targets due to the interference from past trials when the VM targets had served as foils. Both models would predict excellent performance for the CM foils compared to VM foils: the FAM model predicts excellent performance for the CM foils because they are presented very infrequently relative to all other item types; the IR model predicts better performance for CM foils than VM foils because the CM foils are consistently mapped to the "new" responses during training.

Method

Participants

Participants were 51 undergraduates at Indiana University who received credit towards an introductory psychology course requirement.

Stimuli and Apparatus

The stimuli were drawn from a pool of 2,400 unique object images from the website of Talia Konkle from Harvard University. The images were described in Brady, Konkle, Alvarez, and Olivia (2008). Each image subtended a visual angle of approximately 7 degrees and was displayed in the center of a gray background. The experiment was conducted with MATLAB Psychophysics Toolbox (Brainard, 1997) on personal computers.

Procedure

For each subject, 8 stimuli were sampled from the 2400-image set to be the VM-set and another 8 stimuli were sampled to be the CM-set. The 8 CM-set stimuli were divided equally into a CM-target-set and a CM-foil-set. For each trial, the set of to-be-remembered stimuli (memory set) was generated with 2, 4 or 8 items, where half of the items were randomly sampled from the VM-set and the other half were randomly sampled from the CM-target-set. Half the test probes were targets and half were foils. A target test probe was randomly sampled from the memory set; a foil test probe was sampled from the remaining items in the VM-set or the CM-foil-set, with equal probability for each set. Therefore, a CM-target-set item could only serve as a target; a CM-foil-set item could only serve as a foil; and VM-set items switched roles from trial to trial. Because the memory set always consisted of a mixture of CM and VM items in random order, subjects would not be able to tell if the trial was going to be a CM trial or a VM trial until the test probe was presented.

Each subject completed a single session of testing that lasted about 40 minutes in total. The session consisted of 7 blocks with 25 trials for each block. Subjects were instructed to memorize the memory-set items on each trial and indicate if the test probe was a member of the memory set (an old item, or target) or not (a new item, or foil) by pressing a key on the computer keyboard (J=old, F=new). Subjects were not informed of the CM and VM manipulations before testing. On each trial, a fixation point ("*") appeared on the center of the screen for 0.5 seconds to indicate the start of that trial. Then each of the memory-set items was presented in the center of the screen for 1 second, followed by a 0.1 seconds inter-stimulus-interval (ISI). After 1 second of blank screen, another fixation point ("+") appeared for 0.5 seconds, followed by a test probe. The test probe stayed on screen until a key response was registered, after which feedback was provided to

indicate whether or not the response was correct. Subjects were instructed to rest their index fingers on the response keys throughout the experiment and to respond as quickly as possible without making errors.

Results

We considered the first block to be a practice block, so we did not include the data from the first block in our analyses. We also eliminated trials with response time (RT) greater than 3000 ms or less than 180 ms (~3% trials). Finally, we eliminated the data from four outlier subjects who performed significantly worse than the remaining subjects in the group (for VM, overall mean RT greater than 1600 ms or overall accuracy less than 0.6; for CM, overall accuracy less than 0.8).

The main results of the experiment are displayed in the left panels of Figures 2 and 3. In Figure 2, we plot the mean RTs for correct responses as a function of condition (CM vs.VM), set size, and the type of test probe (target vs. foil). The error probabilities are plotted as a function of these variables in Figure 3. The error bars indicate between-subjects standard errors. Because visual inspection indicated that the patterns for targets and foils were quite different from one another, we performed 3 (set size) x 2 (conditions) within-subject ANOVA separately for the targets and foils.

The overall data pattern for foils is highly consistent with previous studies in which CM and VM were trained separately in similar experiments (i.e. Nosofsky, Cox, et al. 2014): Performance for CM is better than for VM, with shorter RTs and lower error rates. The difference is statistically significant for both RT ($F(1, 46) = 151.43$, $p < 0.001$) and the probability of error ($F(1, 46) = 104.67$, $p < 0.001$). Most importantly, there is little set size effect

for CM new items whereas the performance for VM new items declined substantially as the set size increased. The main effect of set size is significant for both RT ($F(2, 92) = 29.31$, $p < 0.001$) and probability of error ($F(2, 29) = 56.18$, $p < 0.001$), but it is mostly driven by the VM foils as the interaction of set size and condition is significant for RT ($F(2, 92) = 17.12$, $p < 0.001$) and probability of error ($F(2, 92) = 63.67$, $p < 0.001$).

In contrast to the dramatic performance difference between CM and VM foils, there was little performance difference between the CM and VM targets when they were mixed within trials. The main result of condition is not significant for either RT ($F(1, 46) < 1$) or the probability of error ($F(1, 46) = 2.59$, $p = 0.14$). In addition, a small but statistically significant set size effect is observed for both CM and VM targets in terms of RT ($F(2, 92) = 20.08$, $p < 0.001$) and probability of error ($F(2, 92) = 6.31$, $p = 0.003$). The important point is that overall performance for the CM targets is far worse under these mixed-list training conditions compared to what has been observed for CM targets in previous studies in which CM and VM training took place separately. Indeed, under these mixed-list conditions, there is little overall difference in performance for the CM vs. VM targets. As explained previously, we expect that such a pattern of results for the CM and VM targets will challenge the IR model.

Model Fitting Results for Experiment.1

Because we are interested in how well the models could account for the qualitative aspects of the data pattern at the group level, we fitted both the FAM and IR models to the averaged group data by minimizing a weighted sum-of-squared deviations (WSSD) criterion. The models were required to simultaneously account for the probability of errors (measured in

proportions) and RT for correct trials (measured in seconds). In particular, the foils were fitted as a function of set size and the targets were fitted as a joint function of set size and lag.[1] To develop a reasonable criterion of fit from all those different data sources, we need to give different weights to the data points based on the different scales of measurement and sample sizes associated with them. For both the current experiment and Experiment 2, we gave (a) the accuracy data twice the weight of the RT data, and (b) data points for foils 4 times the weight of data points for targets (because the individual target data points were broken down by both set size and lag).

Because the CM and VM items were mixed within trials in the present study, we assume that it is unlikely that the subject could adopt different strategies for these different conditions: thus, the only source for different performances is the past history of the items. This assumption is supported by empirical results and quantitative modeling results (Sperling & Dosher, 1986; Strayer and Kramer, 1994b). Therefore, we held most parameters in both models fixed across CM and VM conditions, except for the corresponding LTM parameters in each model. We further constrained the relationship between the LTM parameters based on the theoretical interpretations of each model. Specifically, in the FAM model, we allowed FAM-OLD and FAM-NEW to vary freely with the constraint that FAM-OLD is greater than FAM-NEW for CM items (because CM targets occur with high frequency in the memory sets, whereas CM foils occur only occasionally as test items). For the VM condition, the foils and targets share the same long-term memory familiarity parameter because whether an item serves as target or foil is randomly decided for each trial. In the full version of the IR model, we allowed for the possibility that IR-OLD might be greater than IR-NEW for both CM and VM; the reason is that

old targets also appear in the study lists, which might be viewed as a form of item-OLD response mapping.

The predictions produced from the best-fitting FAM model are plotted in the middle panels (b) of Figures 2 and 3; whereas the predictions from the best-fitting IR model are plotted in the right panels (c) in the same figure. Despite different interpretations of the role of long-term memory, the full versions of both models produced very similar predictions that captured the main pattern of the data. The WSSD of both models are reported in the top panel of Table 1.

The best-fitting parameter values for the FAM model are reported in Table 2. Based on the target-foil asymmetry noted above, a CM foil would be encountered by the subjects with a much lower frequency than a CM target. Consistent with our expectations, the CM FAM-NEW parameter took on a very small value, resulting in short RTs and low error rates for the CM foils. The familiarity parameter for the VM items is relatively large, which leads the model to predict an overall bias toward OLD responses for VM items; the large value of FAM-NEW allows the model to account for the finding that subjects had difficulty in correctly rejecting the VM foils. The fact that the FAM-OLD parameter for CM targets is greater in magnitude than for VM targets is also sensible given the structure of the experiment. In particular, when constructing the memory set, CM targets were more likely to be sampled because half of the memory set came from the CM-target-set (4 items) while the other half was randomly sampled from the full VM-set (8 items). Overall, therefore, individual CM targets occurred more frequently than individual VM items, regardless of whether VM test probes were targets or foils on the current trial. In summary, the FAM model was able to account for the data in this mixed-condition experiment with reasonable parameter values.

The best-fitting parameter values for the IR model are reported in Table 3. The model accounted for superior performance for CM foils through a relatively large setting of the IR-NEW parameter. To account for the dramatically different performances between VM targets and VM foils, the model estimated a large IR-OLD and a near-zero value for IR-NEW. Although the model successfully captured the data pattern, the best fitting parameter values suffered from internal inconsistency: the model essentially predicted that strong associations between CM foils and new responses were learned with relatively little training, while the same amount of training yielded no association between VM foils and new responses. In fact, one could argue that the lack of any item-to-New-response learning makes the IR model indistinguishable from the FAM model in the VM condition: in both cases, the LTM parameters only increase the probability of responding "Old".

To address the problem, different versions of the IR model were fitted to the data to explore whether the near-zero estimate of IR-NEW for the VM foils was a requirement for adequate fits. In one version, we forced IR-NEW to be the same across CM and VM conditions, since the amount of training toward "New response" should be roughly the same for the CM foils and the VM items. Imposing this constraint again resulted in a near-zero estimate for the IR-NEW parameter. While this constrained model yielded essentially the same fit (WWSD = 0.145) as the unconstrained version (WWSD = 0.140), it was completely indistinguishable from the FAM model for both CM and VM conditions. In another constrained version of the IR model, we fixed the amount of IR learning to be the same across "Old" and "New" responses in the VM condition. In other words, IR-OLD would take the same value as IR-NEW for VM items. This constrained version led to substantially worse fits (WWSE = 0.317) than the other versions. Much to my surprise, further inspection of the parameter estimates revealed that even this

version of the model took a near-zero value for IR-NEW in the VM condition. In summary, despite the attempts to produce a sensible value for IR-NEW by constraining it with other parameters, the model persistently set IR-NEW to a near-zero value. The model fitting results can be explained with the following rationale: In the IR model, a relatively large IR-NEW parameter value would increase the probability of responding "New" and most importantly, reduce the set-size effect for the VM foils (because the LTM parameters would reduce the impact from the current list items). Therefore, to account for the large set-size effect observed for the VM foils, the IR model was forced to take a near-zero value for IR-NEW.

Discussion

In Experiment 1, the dramatic differences in performance between CM foils and VM foils are consistent with previous studies where CM and VM manipulations were blocked. This pattern of results is consistent with the hypothesis that CM-based learning changes happen automatically. However, the absence of differences between the CM targets and VM targets has not been reported in previous studies in which CM and VM conditions were separately blocked, suggesting that subjects may develop specialized strategies for CM vs. VM under blocked conditions (see also Strayer & Kramer, 1994a).

Most important, the model-fitting results suggested that the IR model struggled to set parameter values that were consistent with the underlying learning theory. Particularly, the best-fit value for the IR-NEW parameter in the VM condition was near zero, which means the model assumed no item-response learning between VM items and "New" responses. Such a parameter value seems incompatible with the item-response learning theory, at least in its simple form as proposed in the original Instance theory (Logan, 1988). By contrast, the FAM model is able to

successfully account for the data pattern with reasonable parameter values, which suggested that subjects might rely mostly on the familiarity-only process in the mixed condition. Of course, one could try to justify the near-zero parameter value for IR-NEW in VM by assuming the item-response learning in a different form. One possibility was that the target and foil trials compete with one another to some extent in the VM condition, weakening the learned item-response strengths of both. The learning of the NEW response was particularly weakened due to the fact that test probes that had occurred as VM foils on some trials had also appeared as studied memory-set items on numerous other trials. Therefore, a second experiment was carried out to further investigate the precise learning process subjects might engage when different response-mapping manipulations were mixed within trials.

Experiment 2

In the second experiment, we include an "All-New" (AN) manipulation with the CM and VM manipulations, mixed within trials. The AN manipulation refers to a condition where a new set of stimuli is sampled for each trial. Therefore, an AN test probe, target or foil, is never presented in previous trials and therefore would activate no long-term memory traces. The reason for including the AN condition is that LTM parameters are set to zero in both the FAM and IR models, which leads the two models to make dramatically different predictions of performance regarding AN items. If subjects mostly rely on a familiarity-only process to make the Old/New judgement, AN foils would benefit from the lack of long-term memory familiarity and outperform the CM foils. On the other hand, the IR model would predict the opposite pattern given that AN foils receive no training towards "New" responses from long-term memory while CM foils receive consistent training. Furthermore, the two models would generate different predictions for AN targets. The FAM model predicts a substantially worse performance for AN

targets compared to VM targets, since the AN targets activate no familiarity from long-term memory. The IR model would predict the reverse pattern due to the absence of inconsistent training from the past trials for AN targets. We anticipated that the performance in the AN condition could provide some insight into our model selection. However, as will be discussed, the observed patterns of performance with the mixed CM, VM and AN items turned out to challenge both models.

Method

Participants

The participants were 100 undergraduate students at Indiana University; the students received credit towards an introductory psychology course requirement.

Stimuli and Apparatus

The stimuli and apparatus in Experiment 2 were the same as in Experiment 1.

Procedure

The procedure was the same as in Experiment 1 except for the inclusion of "all-new" (AN) items for each trial. The size of the memory set was 3, 6 or 9. For each subject, 6 stimuli were sampled from the 2400-image set to be the VM-set and another 6 stimuli were sampled to be the CM-set (3 images for CM target set, 3 for CM foil set). For each trial, one-third of the memory set came from the VM-set, one-third from CM target set, and one-third of the memory set were

AN items that were never presented on previous trials. On each trial, the AN items were randomly selected from the remaining items in the 2400-image set. On target trials, the test probe was randomly selected from among the items in the memory set. Foil test-probe trials were equally likely to be VM, CM or AN foils; AN foils were randomly selected from the remaining items in the 2400-image set. All other aspects of the procedure were the same as in Experiment 1.

Results

We applied similar data cleaning procedures as in Experiment 1. Data from the first block were not included in the analysis and similar response time cutoffs (greater than 3000 ms or less than 180 ms) were applied to eliminate outlier trials (~ 3% eliminated). Finally, the data from six outlier subjects who performed significantly worse than the remaining subjects in the group (overall mean RT greater than 1600 ms, overall proportion correct less than 0.6 in VM trials; or overall proportion correct less than 0.8 in CM trials) were excluded from the analysis.

The main results of the experiment are displayed in the left panels of Figure 4 and 5. In Figure 4, we plot the mean RTs for correct responses as a function of condition (AN, CM, VM), set size, and type of test probe (target vs. foil). The error probabilities are plotted as a function of these variables in Figure 5. The error bars indicate between-subjects standard errors. Again, because visual inspection indicated that the patterns for targets and foils were quite different from one another, we performed 3 (set size) x 3 (conditions) within-subject ANOVA for targets and foils separately.

Focusing first on only the CM and VM items, the data pattern was highly similar to the one observed in the previous experiment: the performance for CM foils is fast and accurate with

no set size effect, while the performance for VM foils is slow and error prone with a big set size effect; the performance for CM targets and VM targets is very similar and both show a small set size effect. The data provides converging evidence that when presented in mixed conditions, CM targets receive relatively little benefit from consistent-mapping training.

The most interesting result of the current experiment regards the performance for AN items relative to the CM and VM items. We first focus on the foils. For both the accuracy and RT data, the AN foils were qualitatively more similar to the CM foils than the VM foils: in particular, the performance for AN foils was fast and accurate with little set-size effect. For the foils, there was a significant main effect of condition for both RT ($F(2, 186) = 128.72$, $p < 0.001$) and probability of errors ($F(2, 186) = 253.71$, $p < 0.001$). The main effect of set size was also significant for RT ($F(2,186) = 17.80$, $p < 0.001$) and the probability of error ($F(2,186) = 76.90$, $p < 0.001$); however, as in Experiment 1, the set-size effect was mostly driven by the VM foils, with the interaction of set size and condition being significant for both RT ($F(4, 372) = 8.57$, $p < 0.001$) and probability of error ($F(1,46) = 58.78$, $p < 0.001$). However, visual inspection also reveals a quantitative difference between the CM foils and AN foils. Because such a difference is crucial for model selection, we further performed ANOVA tests on the CM and AN data only. There was no significant main effect of condition or set size on probability of error, but that is likely due to the fact that the error rate for both CM and AN is near floor. For the RTs, the main effect of condition was significant ($F(1, 93) = 24.16$, $p<0.001$)[2]. The quantitative difference between the CM and AN foils poses a serious challenge to the FAM model: regardless of how rare CM foils were, they could not be less frequent than AN foils, thus the model cannot predict worse performance for the AN foils compared to the CM foils. The

pattern is more consistent with the IR model where CM foils benefit from more training towards "new" responses.

For both RT and probability of error, performance for CM, VM and AN targets was similar overall; however, AN items did have a small but statistically significant advantage for the error rates ($F(2, 186) = 4.79$, $p = 0.009$. There was also a main effect of set size for both RT ($F(2, 186) = 40.30$, $p < 0.001$) and probability of error ($F(2, 186) = 19.27$, $p < 0.001$). These patterns turned out to be challenging for both the FAM and IR models. Because AN items have the lowest LTM familiarity among the three item types, the FAM model was unable to predict better performance for AN compared to CM and VM targets. For the IR model, although it is natural to predict that performance for AN targets would be better than for VM targets, the model also insisted on predicting an even better performance for the CM targets, which is not supported by the data.

Model Fitting Results for Experiment 2

The model-fitting procedure was similar to the one used in Experiment 1. All parameters except long-term memory parameters were fixed across the AN, CM and VM conditions for both the FAM and IR models. The same constraints were applied to the LTM components for the CM and VM parameters. In addition, the LTM components were fixed to be zero for the AN condition in both the FAM and IR models. The WSSD of both models are reported in Table 1.

Again, we first focus on the FAM model predictions. The best-fitting RT predictions are plotted in panel B of Figure 4, and the best-fitting predictions of the probability of error are plotted in panel B of Figure 5. The best-fitting parameters are reported in Table 4. As discussed in the Results section, we expect the FAM model to have difficulties predicting the pattern of

performance for the CM foils and AN foils. As shown in the figures, the best the FAM model could do was to choose a very small FAM-foil parameter for CM foils and predict the same level of performance for CM and AN foils. The model also had trouble accounting for the quantitatively similarly performances among targets across conditions observed in the data. According to the model, the FAM-OLD parameter increases the probability of responding "Old" for a test probe, and the fact that the FAM-OLD parameter for AN targets was set to be zero means the FAM model would persist predicting a worse performance for the AN targets compared to the CM and VM targets.

The FAM model had additional trouble accounting for the large set-size effect for the VM foils. To predict such a pattern, the model required a large FAM-NEW parameter for VM foils. On the other hand, to predict similar quantitative performance for targets across the CM, VM and AN conditions, the FAM-OLD parameter had to be extremely small for CM *and* VM targets because the FAM-OLD parameter is fixed to be zero for AN. Because FAM-OLD and FAM-NEW are constrained to be equal for VM, the model appears to have compromised by choosing a non-zero but low-magnitude value for FAM-OLD and FAM-NEW for the VM items, and thereby underestimates the magnitude of the set-size effect for the VM foils.

The best-fitting predictions from the IR model are plotted in panel C of Figures 4 and 5. The best fitting parameters are reported in Table 5. Overall, the IR model provided a much better fit to the data than the FAM model. For the foils, the IR model successfully captured the qualitative relationship holding among the CM, VM and AN items. The model managed to predict better performance for CM foils by allowing a relatively large CM IR-new parameter; the performance for VM foils suffers from a relatively large IR-OLD parameter; and the AN foils receive neither the benefit nor harm from the long-term memory traces. However, the model fell

short of predicting the data pattern for the targets: while it managed to predict similar quantitative performance between the AN and VM targets by choosing a large VM IR-OLD parameter and a small VM IR-NEW parameter, the model struggled to predict similar performance for CM targets since the CM IR-OLD parameter is constrained to be larger than the CM IR-NEW parameter (given the CM targets potentially receive additional item-old association training through their presentations on the study lists). In addition, the IR model also had trouble accounting producing the large set-size effect for VM foils. In the IR model, a large IR-NEW parameter means better performance for foils and a *smaller* set-size effect, since the LTM parameter values were not influenced by the current list. To account for the patterns for foils, the model requires a larger IR-NEW for the AN foils than for the VM foils. However, the IR-NEW parameter for AN foils was set to zero while the IR-NEW parameter for VM foils was constrained to be non-zero. Therefore, the model overestimates the set-size effect for AN foils but underestimates the set-size effect for VM foils^3.

Discussion

In Experiment 2, we replicated the same overall data pattern for CM and VM items from Experiment 1. The two experiments provided converging evidence suggesting that the dramatic performance differences between CM targets and VM targets when trained in blocked fashion were largely reduced and eliminated when they were mixed within trials.

Contrary to the modeling fitting results for Experiment 1 where both the IR model and the FAM model were able to capture the data pattern of mixed CM and VM items (albeit with unusual parameter settings for the IR model), both models were severely challenged by the addition of AN items in the present mixed condition. Specifically, the performance for AN foils

was not as good as the performance for the CM foils. This result strongly challenges the FAM model because CM foils must have at least as much LTM familiarity as AN foils. On the other hand, performance on the AN foils was nearly invariant with set size. This result challenges the IR model because the absence of the training from long-term memory means the performance is more sensitive to the current list – and the model is forced to predict large set-size effects under such conditions. In summary, the data patterns in Experiment 2 pose a significant challenge to both models.

General Discussion

Summary. The goal of the present work was to investigate whether a common learning process, particularly an item-response learning process (IR model), could account for the diverse patterns of performance observed in probe-recognition memory search tasks in cases in which the history of assignment of items as targets and foils was manipulated. To test the theory, we conducted two experiments where we mixed items that were consistently mapped (CM) to a particular response (old vs. new) with items that either had no past history (AN) or were randomly assigned to different responses across trials (VM). In Experiment 1, we found that the dramatic differences between CM targets and VM targets that were reported in numerous past studies done with blocked manipulation were eliminated in the present study where we mixed CM items with VM items within trials. On the other hand, the differences between CM foils and VM foils persisted in the mixed condition. In Experiment 2, where CM items were mixed with both VM and AN items, we found the same patterns between CM and VM items were replicated from the previous experiment. Moreover, the patterns of performance for the AN items relative to the CM and VM items provided further constraints for diagnosing the nature of the learning and memory processes that are involved.

The exemplar-based random-walk model (EBRW; Nosofsky and Palmiri, 1997) was modified to explicitly model how past history affects the task performance in terms of accuracy and response time. Both an item-response-learning (IR) version of the model and an item-familarity (FAM) version were fitted to the data of both Experiments 1 and 2. Although the IR version of the model had provided excellent accounts of performance in past studies (e.g., Nosofsky, Cao et al. 2014; Cao et al. 2018), it failed to account for the performance patterns in the present experiments. In Experiment 1, it was able to fit the data only with certain parameter settings that were inconsistent with the form of item-response-learning process that underlies the theory. And in Experiment 2 it failed to capture the behavior patterns regardless of the settings of its free parameters. The FAM model accounted well for the data pattern in Experiment 1, but then failed to capture the data pattern in Experiment 2. Overall, the model fitting results challenge the idea that performance can be captured by unitary learning and memory processes based solely on either item-response-learning or familiarity-based mechanisms.

Mixed Condition Data Patterns

Although the focus of this study was on testing certain extant versions of the EBRW probe-recognition model, there are some salient empirical patterns that are likely to have bearing for numerous other models as well. For performance in the CM, VM and AN conditions, the very robust patterns that were established in various previously reported experiments with blocked training changed dramatically in the present studies when the items from different manipulations were mixed within trials. Perhaps the most puzzling changes occurred for the AN items: when trained in blocked fashion, one typically observes a substantial set-size effect for both AN targets and AN foils (e.g., Nosofsky, Cox, et al., 2014); but when mixed with CM and

VM items, the performance for AN items improved substantially and, most importantly, there was no sign of a set-size effect for the AN foils. A standard familiarity-based process model could explain the behavior pattern for AN items in blocked conditions (for details, see Nosofsky, Cox, et al. 2014), but not under the current mixed-list conditions. Note that the lack of a set-size effect could not simply be attributed to a ceiling effect in performance (because the performance for CM foils was still better), nor could it be explained by a switch to an item-response learning process, as was described in the model-fitting section for Experiment 2. Such results indicate that patterns of probe-recognition performance are dramatically influenced by the specific manipulations of the experiment. These manipulations may have strong influences on rates of learning for different item types and may also lead to major adjustments in response strategy.

Dual-process Models

The model-fitting results strongly suggest that it would be difficult for a single-process learning model to account for the data pattern in the present studies. It appears that some type of dual-process model may be required. In our present modeling approach, retrieved examples from short-term memory and long-term memory influenced evidence accumulation in a single random-walk process for making decisions. An alternative idea is that there may be two separate processes governed by short-term retrieval and long-term retrieval and that these operate in parallel, with the process that completes first leading to the old-new recognition decision. It seems likely that such a model would predict well the fast and accurate performance observed for CM items (e.g., Logan, 1988). However, such a model would likely have difficulty accounting for other aspects of our results. For VM, a retrieved response from long-term memory could win the race, but such an occurrence would result in fast and near-chance

performance. Alternatively, VM retrieval from long-term memory never wins due to the inconsistent response mapping, in which case the performance completely relies on information in the current trial. In that case, however, the model predicts the same performance for VM and AN items. Neither prediction is supported by the data, where performance for VM items is often slower than performance for AN items. The data pattern seems to require a more flexible architecture for the interaction between the two systems.

An alternative dual-process architecture is proposed in Mewhort and Johns (2005). The researchers argued against the popular assumption that a "new" response is simply a result of insufficient evidence towards an "Old" response. Instead, subjects accumulate evidence toward "New" responses and evaluate that evidence separately from the evidence for "Old" decisions. The dramatically diverse behavior patterns observed for foils across our different training conditions could potentially be explained with these types of separate decision systems: Subjects first accumulate and evaluate evidence towards "new" responses at a global level (combining both the long-term memory and the short-term memory with no special emphasis on the current list), either through accumulating traces with new response labels or through lack of overall familiarity. A sufficient amount of such evidence would result in a quick "New" response. If a test probe fails to lead to a "New" response at this global stage, then it will be compared in a more focused manner to the current list items, where a match leads to an "Old" response and a non-match to a "New" response. The "New" response for foils in CM and AN is likely to occur at the global level, leading to RTs that are largely invariant with the set size of the current list. On the other hand, the "New" response for VM foils is likely to be made by comparing to current list items, hence a strong set-size effect emerges. I plan to explore different models based on

these ideas that could potentially account for the full range of data patterns observed in both blocked or mixed designs in future research.

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes1. In *Psychology of learning and motivation* (Vol. 2, pp. 89-195). Academic Press.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, USA, 105*, 14325– 14329.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Cao, R., Nosofsky, R. M., & Shiffrin, R. M. (2017). The development of automaticity in short-term memory search: Item-response learning and category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43(5),* 669.

Cao, R., Shiffrin, R. M., & Nosofsky, R. M. (2018). Item frequency in probe-recognition memory search: Converging evidence for a role of item-response learning. *Memory & cognition*, *46*(3), 450-463.

Cheng, P. W. (1985). Restructuring versus automaticity: Alternative accounts of skill acquisition. *Psychological review, 92*, 414-423.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive psychology*, *6*(2), 293-323.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*, 492–527.

Logan, G. D. (1990). Repetition priming and automaticity: Common underlying mechanisms?. *Cognitive Psychology*, *22*(1), 1-35.

Logan, G. D., & Stadler, M. A. (1991). Mechanisms of performance improvement in consistent mapping memory search: Automaticity or strategy shift?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(3), 478.

McElree, B., & Dosher, B. A. (1989). Serial position and set size in short-term memory: Time course of recognition. *Journal of Experimental Psychology: General, 18*, 346–373.

Mewhort, D. J. K., & Johns, E. E. (2005). Sharpening the echo: An iterative-resonance model for short-term recognition memory. *Memory*, *13*(3-4), 300-307.

Monsell, S. (1978). Recency, immediate recognition memory, and reaction time. *Cognitive Psychology, 10*, 465–501.

Nosofsky, R. M., Cao, R., Cox, G. E., & Shiffrin R. M. (2014). Familiarity and categorization processes in memory search. *Cognitive Psychology, 75*, 97-129.

Nosofsky, R. M., Cox, G. E., Cao, R., & Shiffrin, R. M. (2014). An exemplar-familiarity model predicts short-term and long-term probe recognition across diverse forms of memory search. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1524.

Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review, 188*, 280–315

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*, 266–300

Nosofsky, R. M., & Stanton, R. D. (2005). Speeded classification in a probabilistic category structure: Contrasting exemplar-retrieval, decision-boundary, and prototype models. *Journal of Experimental Psychology: Human Perception and Performance, 31(3)*, 608-629.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108.

Schneider, W., & Chein, J. M. (2003). Controlled & automatic processing: behavior, theory, and biological mechanisms. *Cognitive science*, *27*(3), 525-559.

Schneider, W., & Fisk, A. D. (1982). Degree of consistent training: Improvements in search performance and automatic process development. *Perception & Psychophysics, 31(2),* 160–168.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84*, 1–66

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84*, 127–190.

Strayer, D. L., & Kramer, A. F. (1994a). Strategies and automaticity: I. Basic findings and conceptual framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(2)*, 318–341.

Strayer, D. L., & Kramer, A. F. (1994b). Strategies and automaticity: II. Dynamic aspects of strategy adjustment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(2), 342.

Sternberg, S. (1966, August 5). High-speed scanning in human memory. *Science, 153*, 652– 654. http://dx.doi.org/10.1126/science.153.3736.652

Sternberg, S. (2016). In defence of high-speed memory scanning. *The Quarterly Journal of Experimental Psychology*, *69*(10), 2020-2075.

Footnotes

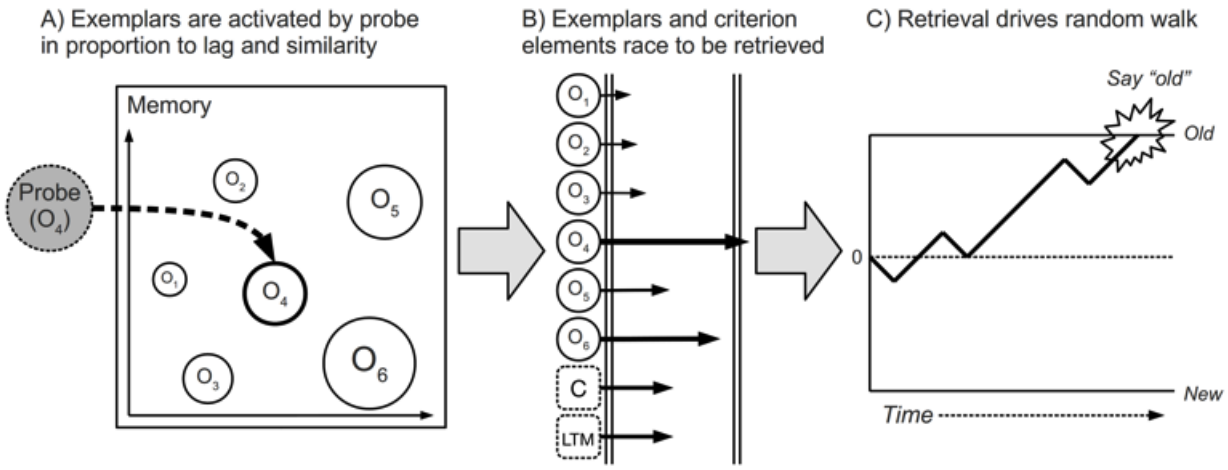
1. Although sometimes noisy due to small sample sizes, the more detailed patterns of set-size by lag data closely resembled the ones we have reported in previous studies. In particular, performance declined with increases in lag; furthermore, once we conditioned on lag, there was little remaining difference in old-item performance across the different set-size conditions. These same patterns also held for our Experiment-2 data. These more detailed data from both Experiments 1 and 2 are illustrated in figures in the appendix, along with the more detailed predictions from the FAM and IR models.

2. In the 2 (conditions) x 3 (set size) ANOVA test for CM and VM foils, the main effect of set-size was also statistically significant ($F(2, 186) = 3.60$, $p = 0.029$). However, the performance differences across the different set sizes were quantitatively small: the difference between the largest mean RT and smallest mean RT for AN foils was 50.3 ms; the difference between the largest mean RT and smallest mean RT for CM foils was 12.9 ms.

3. Theoretically, IR model could set the VM IR-NEW parameter to be near-zero (0.001, the minimum value for the parameter), which could result in a more pronounced set-size effect for VM foils. We tried to fix the VM IR-NEW value to 0.001 during the parameter search, which resulted in an overall quantitatively worse fit (WWSE = 0.30).

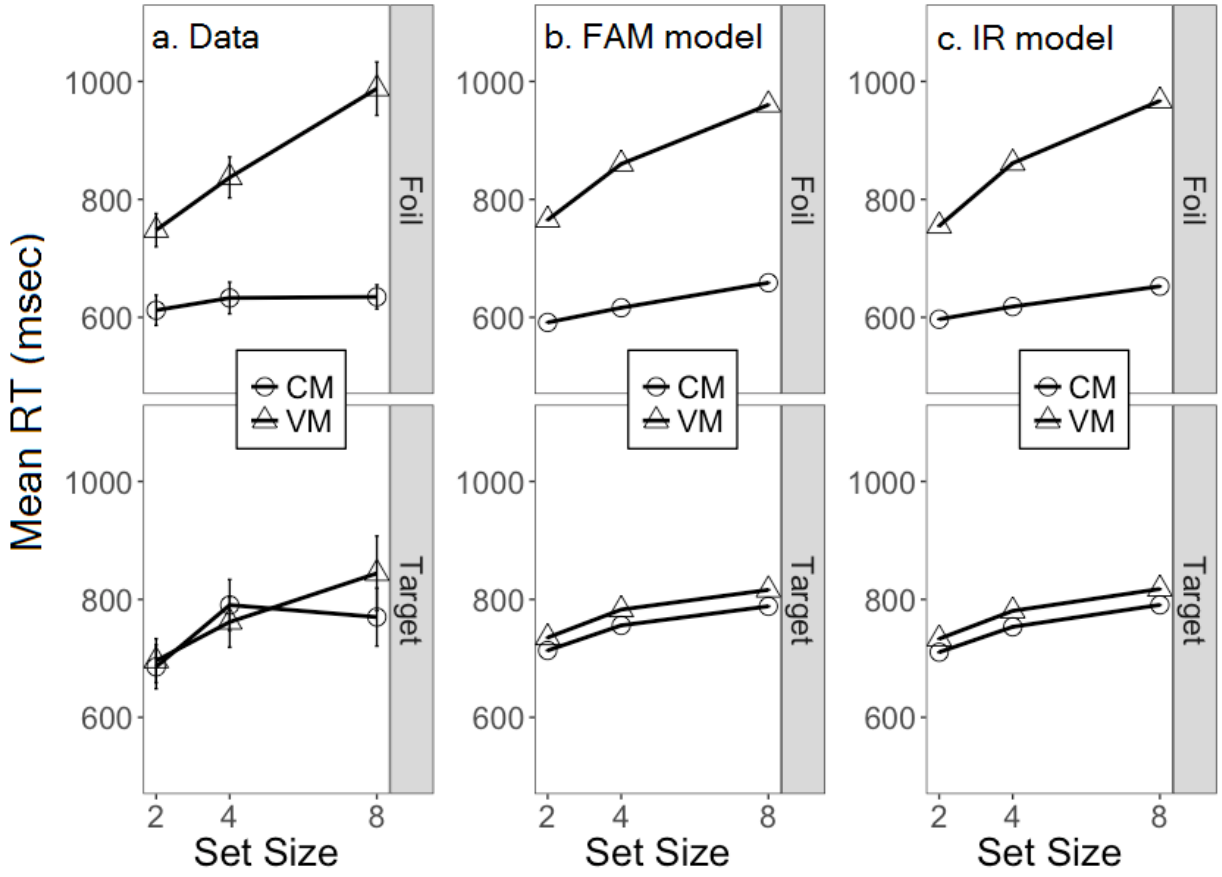
Figures

Figure 1



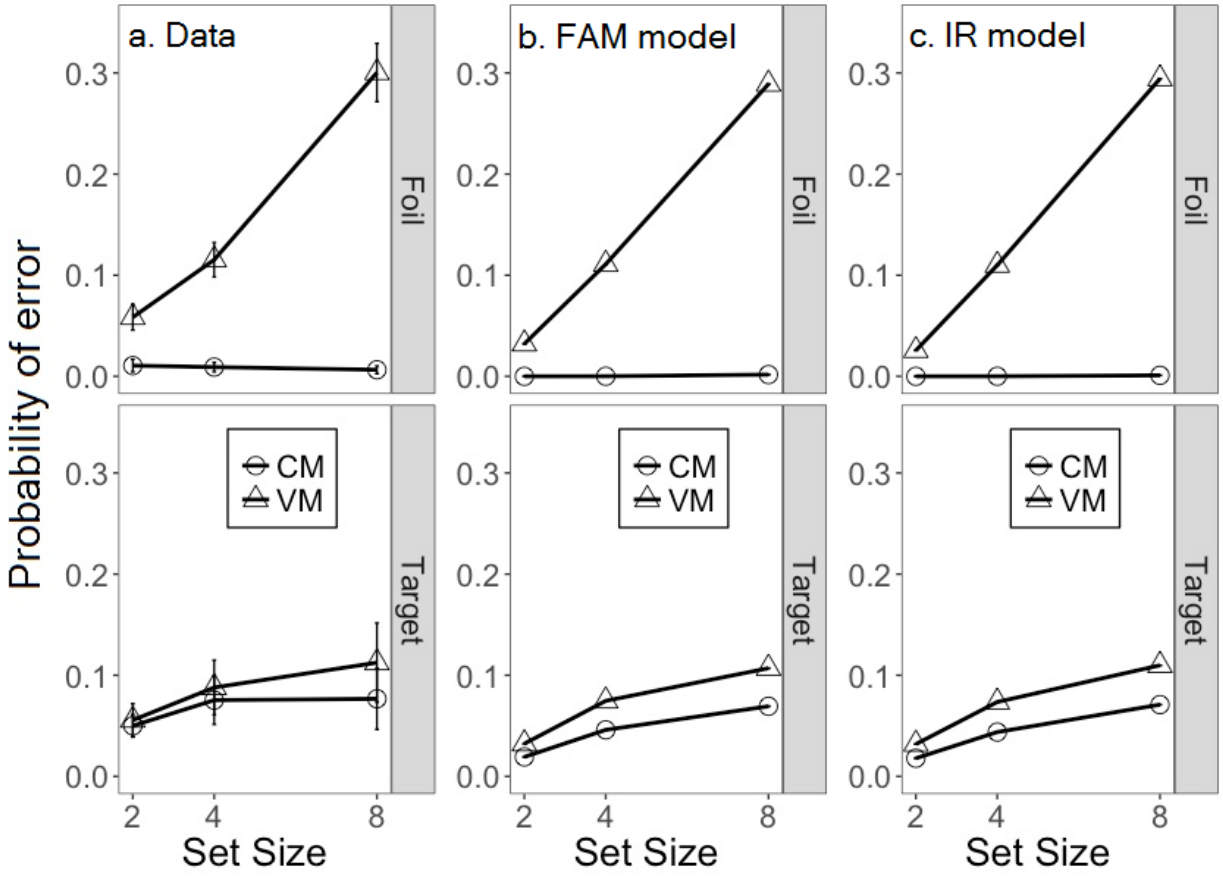
Schematic illustration of the application of the exemplar-based random-walk model to the short-term probe-recognition task. Note: O_k is the old item on the current study list that is presented in serial-position k.

Figure 2



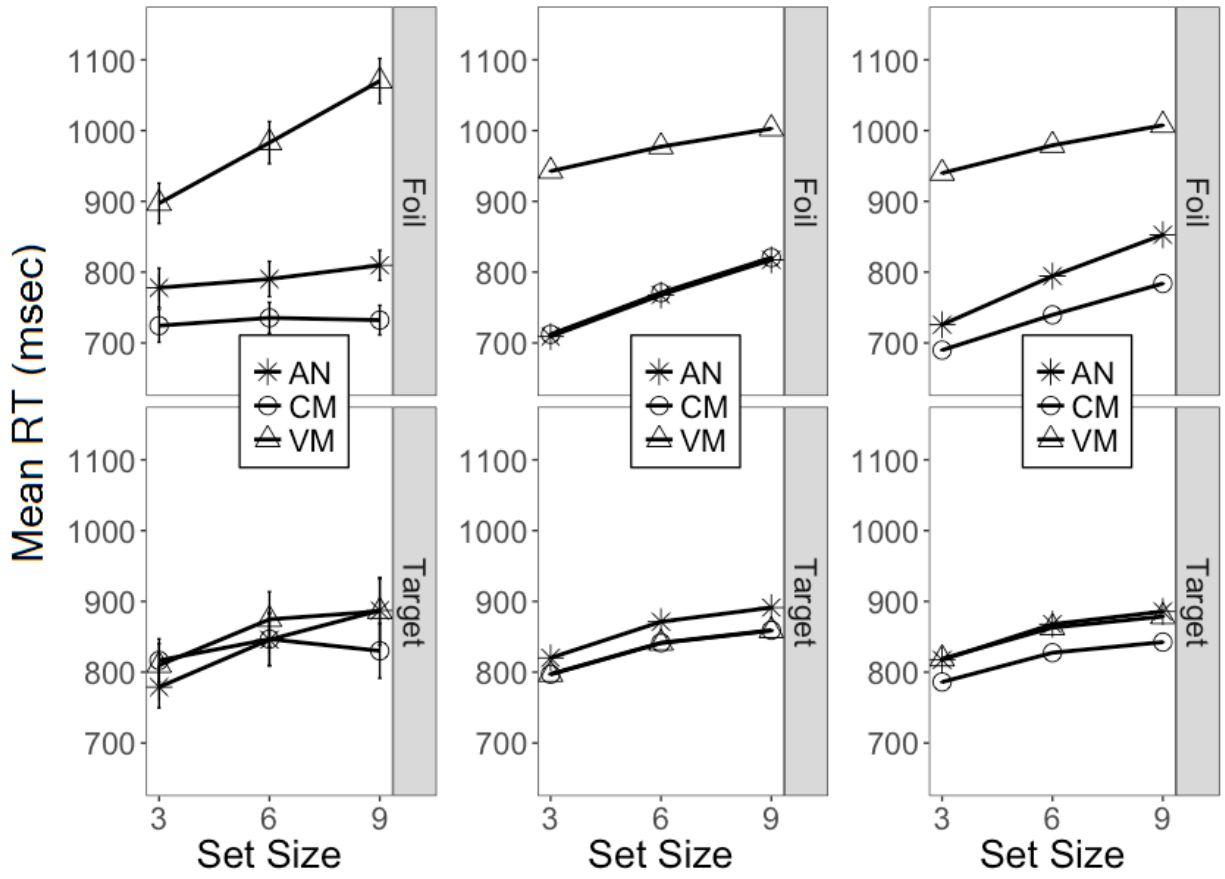
Mean correct response times in Experiment 1 plotted as a function of conditions (CM, VM), set size, test-probe type (target, foil). Left panel = observed data, middle panel = predictions from full version of familiarity model, right panel = predictions from full version of item-response-learning model. Error bars show the between-subject standard-error of the mean.

Figure 3



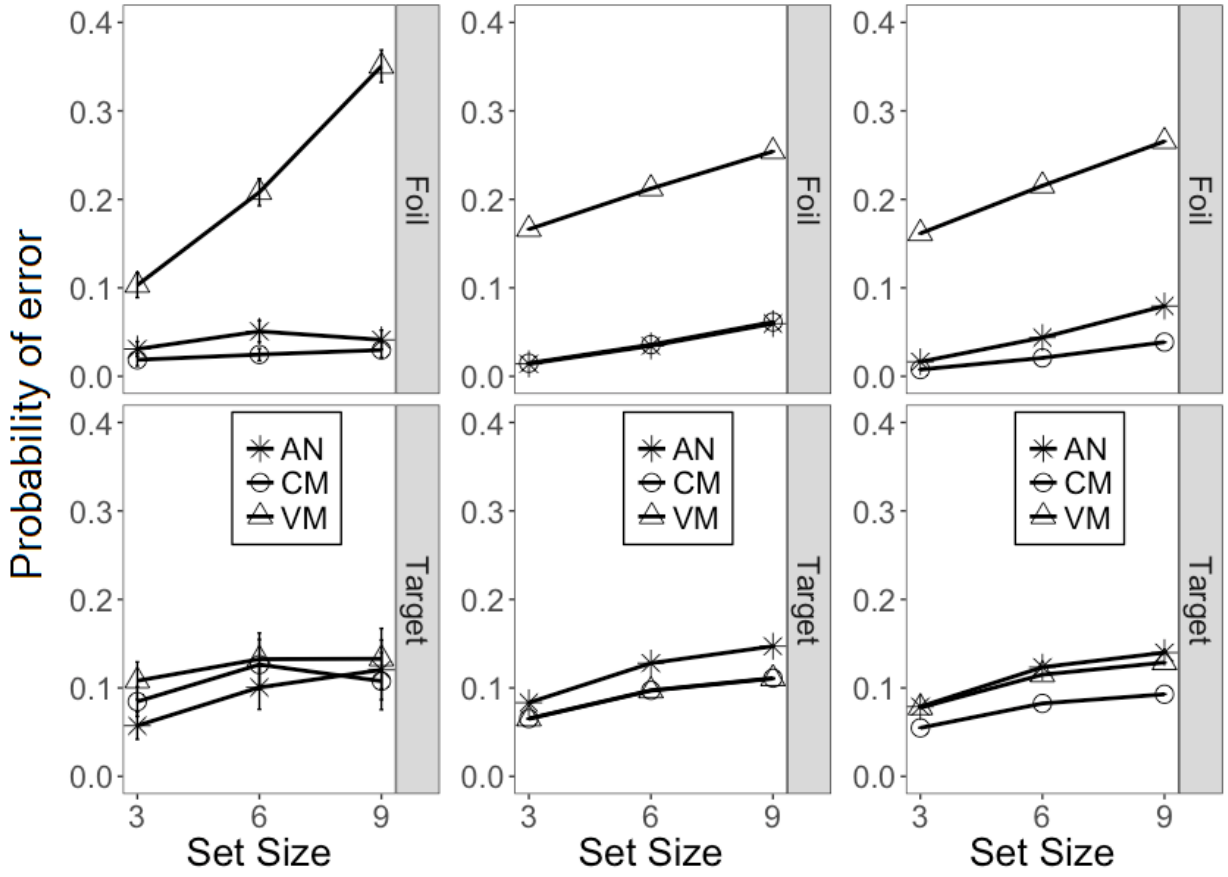
Probability of errors in Experiment 1 plotted as a function of conditions (CM, VM), set size, test-probe type (target, foil). Left panel = observed data, middle panel = predictions from full version of familiarity model, right panel = predictions from full version of item-response-learning model. Error bars show the between-subject standard-error of the mean.

Figure 4



Mean correct response times in Experiment 2 plotted as a function of conditions (CM, VM, AN), set size, test-probe type (target, foil). Left panel = observed data, middle panel = predictions from full version of familiarity model, right panel = predictions from full version of item-response-learning model. Error bars show the between-subject standard-error of the mean.

Figure 5



Probability of errors in Experiment 2 plotted as a function of conditions (CM, VM, AN), set size, test-probe type (target, foil). Left panel = observed data, middle panel = predictions from full version of familiarity model, right panel = predictions from full version of item-response-learning model. Error bars show the between-subject standard-error of the mean.

Tables

Table 1. Weighted Sum of Squared Deviation (WSSD) Fits of the FAM and IR Models to the Mean Correct RTs and Error-Probability Data in Experiment 1 and Experiment 2.

Experiment 1			
Model	**VM**	**CM**	**Total**
FAM model	0.08	0.06	0.14
IR model	0.08	0.06	0.14

Experiment 2				
Model	**VM**	**CM**	**AN**	**Total**
FAM model	0.12	0.12	0.09	0.33
IR model	0.08	0.11	0.09	0.28

*Table 2***.** The best-fitting parameter values for the full FAM model in Experiment 1.

Parameters (FAM)	CM	VM
α	0.11	-
β	0.83	-
u	0.98	-
v	0.02	-
s	0.15	-
FAM-OLD*	0.51	0.45
FAM-NEW	0.001	**
A_old	9.27	-
B_new	11.78	-
κ	4.19	-
Tr	507.78	-

Note. FAM = item-familiarity model, CM = consistent-mapping, VM = varied-mapping. Parameter values replaced with " – " were set equal to one another across the CM and VM conditions; Parameters value replaced with "**" were constrained to be the same value as the parameter immediately above; Parameters marked with "*" were constrained to be greater than or equal to the parameter immediately below. α = power-decay asymptote, β = power-decay rate, *u* = criterion intercept, *v* = criterion slope, *s* = similarity, A_old = old threshold, B_new = new threshold, κ = scale, Tr = residual time. See text for an explanation of the LTM parameters

Table 3. The best-fitting parameter values for the full IR model in Experiment 1.

Parameters (IR)	CM	VM
α	0.13	-
β	0.87	-
u	0.86	-
v	0.03	-
s	0.17	-
IR-OLD*	0.37	0.31
IR-NEW	0.37	0.001
A_old	2.33	-
B_new	2.58	-
κ	4.79	-
Tr	508.98	-

Note. IR = item-response learning model, CM = consistent-mapping, VM = varied-mapping. Parameter values replaced with " – " were set equal to one another across the CM and VM conditions; Parameters marked with "*" were constrained to be greater than or equal to the parameter immediately below. α = power-decay asymptote, β = power-decay rate, *u* = criterion intercept, *v* = criterion slope, *s* = similarity, A_old = old threshold, B_new = new threshold, κ = scale, Tr = residual time. See text for an explanation of the LTM parameters

Table 4. The best-fitting parameter values for the full FAM model in Experiment 2.

Parameters (FAM)	CM	VM	AN
α	0.43	-	-
β	1.50	-	-
u	0.29	-	-
v	0.001	-	-
s	0.014	-	-
FAM-OLD*	0.095	0.096	(0)
FAM-NEW	0.001	**	(0)
A_old	2.10	-	-
B_new	2.30	-	-
κ	175.1	-	-
Tr	197.19	-	-

Note. FAM = item-familiarity model, CM = consistent-mapping, VM = varied-mapping. Parameter values replaced with " – " were set equal to one another across the CM and VM conditions; Parameters value replaced with "**" were constrained to be the same value as the parameter immediately above; Parameters value incased with "()" were constrained were fixed to particular values; Parameters marked with "*" were constrained to be greater than or equal to the parameter immediately below. α = power-decay asymptote, β = power-decay rate, *u* = criterion intercept, *v* = criterion slope, *s* = similarity, A_old = old threshold, B_new = new threshold, κ = scale, Tr = residual time. See text for an explanation of the LTM parameters

Table 5. The best-fitting parameter values for the full IR model in Experiment 2.

Parameters (IR)	CM	VM	AN
α	0.41	-	-
β	1.49	-	-
u	0.31	-	-
v	0.001	-	-
s	0.02	-	-
IR-OLD	0.124	0.124	(0)
IR-NEW	0.123	0.052	(0)
A_old	2.33	-	-
B_new	2.58	-	-
κ	128.47	-	-
Tr	271.14	-	-

Note. IR = item-response learning model, CM = consistent-mapping, VM = varied-mapping. Parameter values replaced with " – " were set equal to one another across the CM and VM conditions; Parameters value incased with "()" were constrained were fixed to particular values; Parameters marked with "*" were constrained to be greater than or equal to the parameter immediately below. α = power-decay asymptote, β = power-decay rate, *u* = criterion intercept, *v* = criterion slope, *s* = similarity, A_old = old threshold, B_new = new threshold, κ = scale, Tr = residual time. See text for an explanation of the LTM parameters

Appendix

Observed target-item data and predictions from the FAM model and IR model plotted as a joint function of set-size, lag, and conditions. Figure A1 plots mean correct RT in Experiment 1; Figure A2 plots the probability of error in Experiment 1; Figure A3 plots mean correct RT in Experiment 2; and Figure A4 plots the probability of error in Experiment 2.

Figure A1

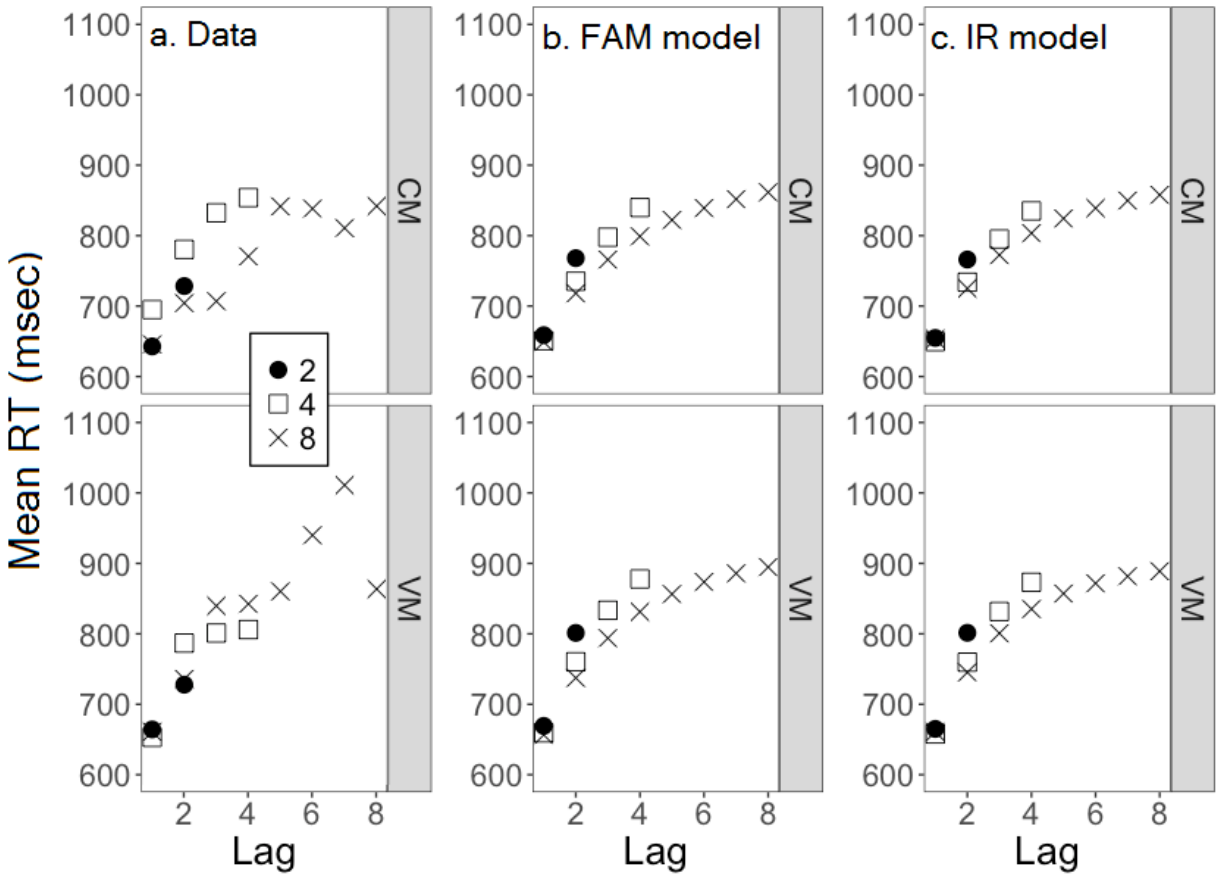


Figure A2

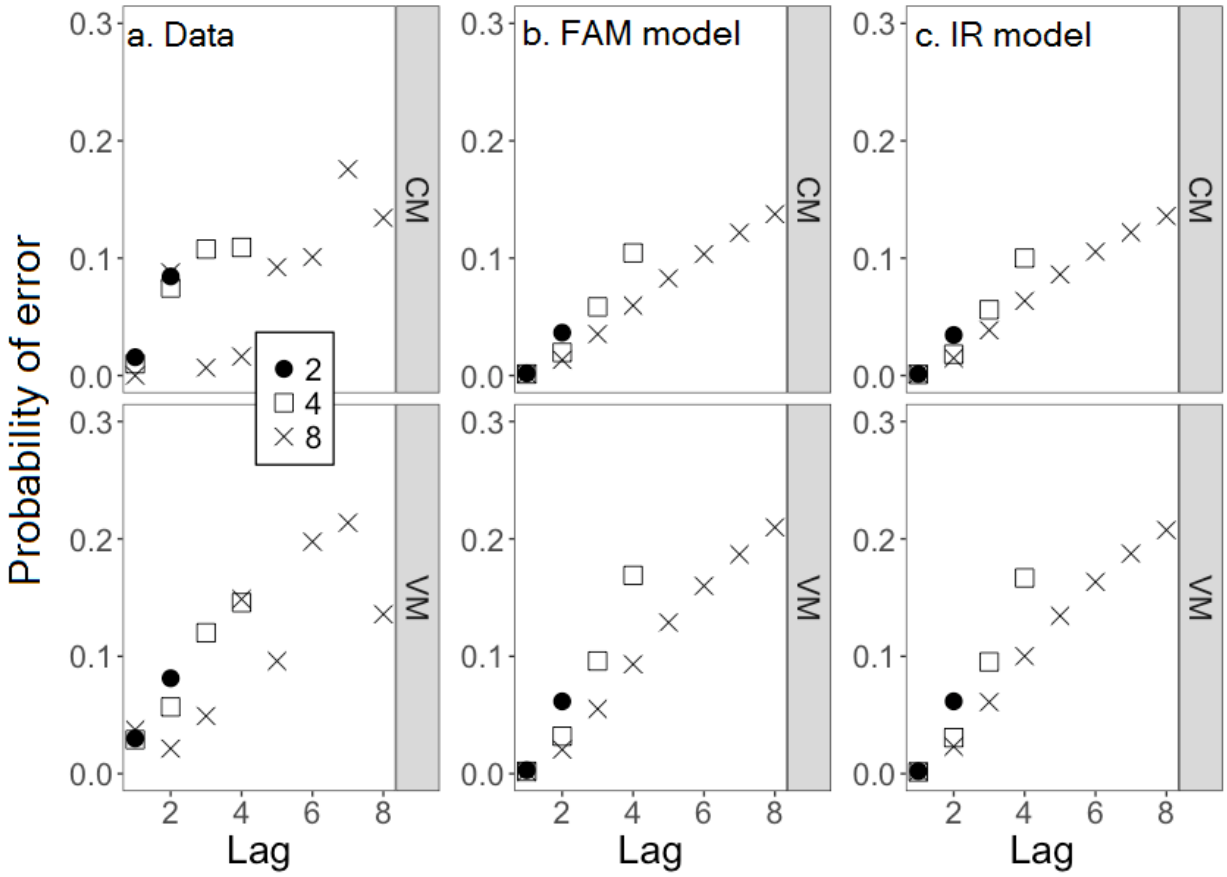

a. Data
b. FAM model
c. IR model
CM
VM
Probability of error
Lag
2
4
8
0.0
0.1
0.2
0.3
2
4
6
8

Figure A3

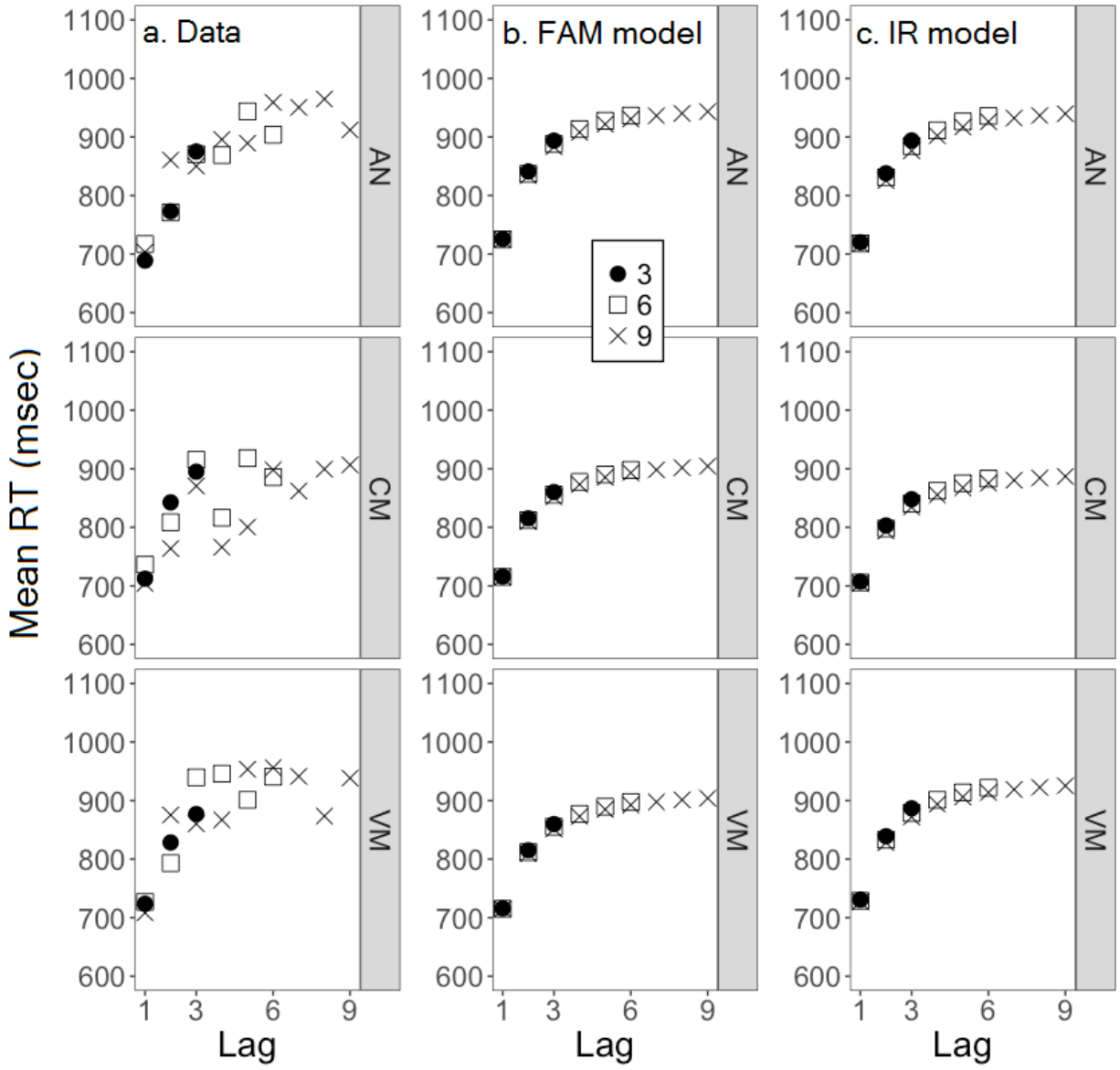
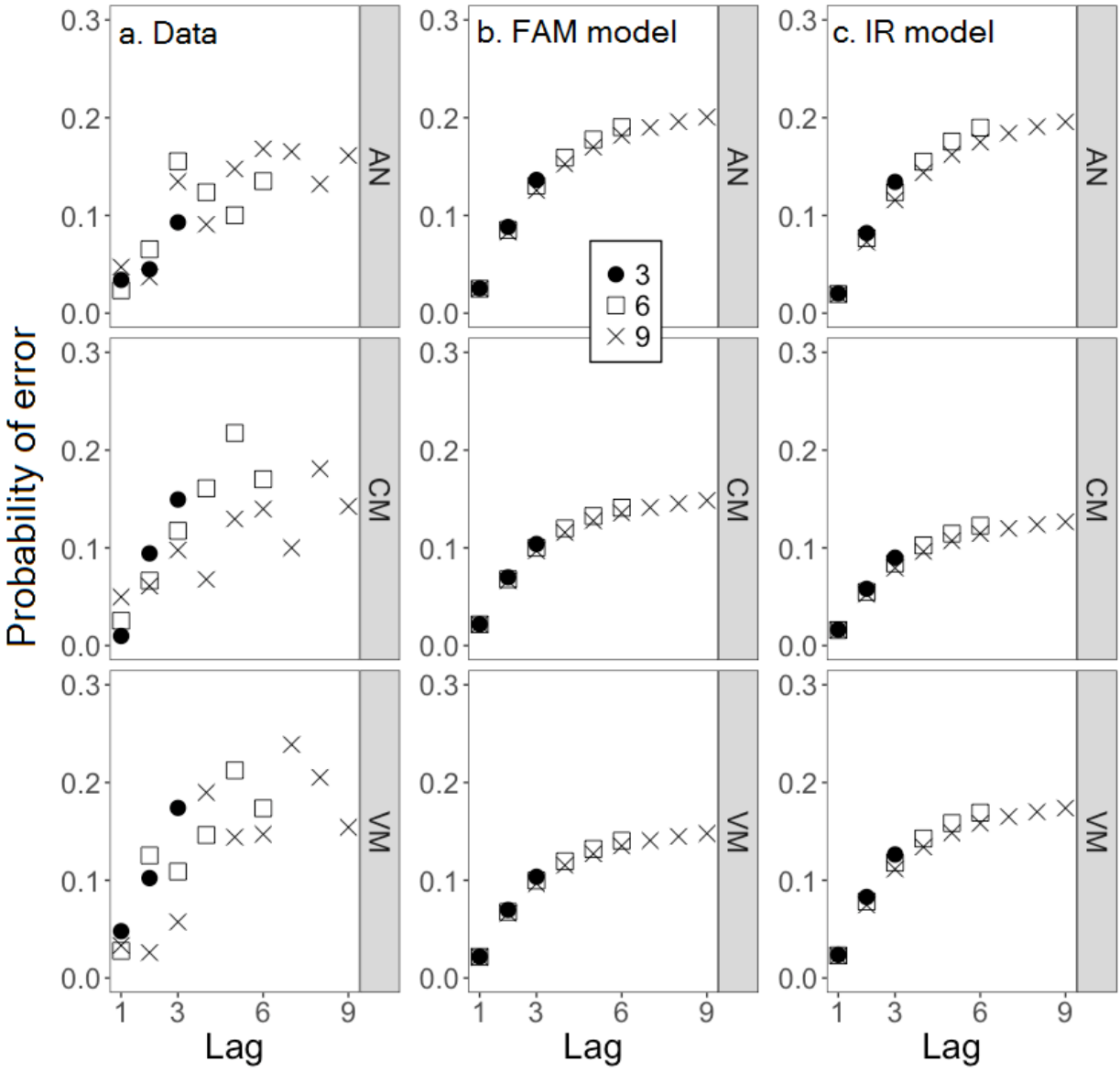

a. Data
b. FAM model
c. IR model
AN
CM
VM
Mean RT (msec)
Lag
3
6
9
1100
1000
900
800
700
600
1
3
6
9

Figure A4


a. Data
b. FAM model
c. IR model
AN
CM
VM
Probability of error
Lag
0.0
0.1
0.2
0.3
1
3
6
9
3
6
9

Tracking the Development of Automaticity in Memory Search with Human Electrophysiology

For many years researchers have studied the nature of different forms of memory retrieval. Sperling (1960) demonstrated highly accurate retrieval from a short-lived memory termed the visual icon, with subsequent less-accurate retrieval from longer lasting memory stores that had lower capacity. Schneider and Shiffrin (1977) and Shiffrin and Schneider (1977) explored the effects of learning on memory retrieval by varying the way that stimulus-response relations are experienced: They trained using either *varied mapping* (VM) in which the binary responses to a given stimulus varied throughout training, or *consistent mapping* (CM) in which the same response was always assigned to a given stimulus throughout training. VM and CM produced marked differences in performance, and this was interpreted as changes in the learning of automatic responses, causing changes in attention and memory retrieval. The memory and visual search paradigms they used were associated with large differences in perceived effort, CM coming to seem relatively effortless while VM remained highly effortful throughout training. Logan (1988) emphasized the role of memory retrieval in studies of the learning of alphabet-arithmetic, showing a switch from effortful algorithmic calculation of answers to relatively effortless and fast memory-based retrieval after consistent training.

In recent years, Cao, Shiffrin and Nosofsky (2018; Nosofsky, Cao, et al., 2014) have used VM and CM training to explore in greater detail their role in storage and retrieval in short-term probe-recognition tasks. In their usual paradigm each trial involves presentation of a short list of to-be-remembered items (usually pictures). The study list is followed by a test probe. Subjects respond "old" if the test probe is an item that appeared on the presented list ("targets"); and

respond “new” if the test probe is an item that did not appear on the list (“foils”). Note that both targets and foils may have occurred as either study items or test probes on numerous previous lists. Consistent with earlier findings, there were marked differences in performance due to VM versus CM training: VM performance did not improve with training and produced large set size and serial position effects. CM performance showed rapid improvement with training and any set size or serial position effects were greatly reduced. These effects were observed in both accuracy and response times.

To explain the findings of VM and CM training on short-term probe recognition, Cao et al. (2018) used a variant of the “Exemplar Based Random Walk” (EBRW) of Nosofsky and Palmeri (1997). In their modeling approach, study and test-probe exemplars presented on previous trials might be retrieved along with current-list exemplars in driving observers’ old-new recognition judgments. VM training caused storage of previous-trial exemplars with roughly equal numbers of old and new responses, which would lead to interference in making old-new judgments for current-list items. CM training produced storage of previous-trial exemplars with consistent responses, which would lead to facilitation of performance. Therefore, in VM, an observer would attempt to limit retrieval to current-list items, placing the emphasis on short-term retrieval. But in CM, an observer can rely on long-term memory retrieval of the consistently mapped exemplar-response pairs established throughout training.

Woodman and colleagues (e.g. Carlisle, Arita, Pardo & Woodman, 2011; Woodman, Carlisle & Reinhart, 2013) used a visual-search paradigm in conjunction with EEG measurements to illuminate possibly different forms of retrieval across VM and CM conditions. In their paradigm a single display of a small number of simple visual stimuli (Landolt C’s in various orientations) were presented on both sides of fixation, the items to be remembered

(targets) being indicated by the color of the stimuli on one side. After the presentation of the cue, the subject was asked to maintain the targets during a delay period, followed by a display of a ring of Landolt C's, which the subject searched for the presence of the studied targets. The researchers observed that, during the delay period, there was a significantly stronger activation from lateral-occipital electrodes on the contralateral side vs. the ipsilateral side of the to-be-remembered item. The difference between contralateral side and ipsilateral side is termed *contralateral delay activity* (CDA; for a recent comprehensive review of the CDA as a neural measure of visual working memory, see Luria et al., 2016; for early evidence, see Vogel & Machizawa, 2004). This CDA signal is stronger when more stimuli must be maintained on one side (e.g. one versus two Landolt C's). Following earlier work, Woodman et al. (2013) suggested the CDA signal provides a measure of the active maintenance of items in short-term visual memory and that it is subjected to top-down attention modulation. In a VM situation in which the to-be-remembered targets varied from trial to trial the magnitude of the CDA remained unchanged from trial to trial. However, in a condition in which the same target repeated for 7 consecutive trials (a form of "local" CM training involving a single item), the magnitude of the CDA signal dropped at each presentation (in another condition it disappeared when subjects searched the same item for an entire session). Carlisle et al. (2011) suggested that in CM long-term memory for the target item could be used, reducing the need to maintain items in visual working memory, thereby reducing the CDA. In VM, no learning could occur, so working memory maintenance was required on every trial.

These findings and interpretations occurred in a task that differed in many ways from our short-term probe-recognition studies described earlier, including the simultaneous versus sequential presentation of the to-be-remembered stimuli; the simplicity of the stimuli (Landolt

C's versus pictures of objects); and the number of different stimuli used in the study. For example, in Woodman's CM paradigm, there was only a single target that repeated consecutively across trials, whereas in traditional CM memory-search studies the test probe is drawn from a large set of stimuli and the specific test probe changes across trials. In this study we therefore returned to a variant of the short-term probe-recognition paradigm, but collected the EEG measures that Woodman found diagnostic in his task. We hoped that the EEG measures could be used to help interpret the differences between VM and CM training.

Participants were presented with short lists of to-be-remembered pictorial stimuli (see Figure 1). The pictures were presented successively on both sides of fixation, one picture on each side. The side of each to-be-remembered item could vary from one visual frame to the next and was indicated by the color of an outline square. This varying presentation-side procedure was adopted in order to reduce subjects' urge to move eyes from fixation, but our interest is on the trials with target stimuli all on one side. In one condition we used VM training for 100 trials (target and foil pictures exchanged roles from trial to trial) and in another condition we used CM training for 100 trials (target and foil pictures maintained their roles over all trials). Our primary interests were in the behavioral accuracy and response time data, and in the CDA measures after each successive item was presented in the study list. (As we will describe, however, we also examined another EEG measure based on Alpha power suppression.) The primary question was whether the CDA signal would be stronger in VM than in CM, providing converging evidence for the differential reliance on STM vs. LTM across these different conditions of memory search. In addition, we were interested in exploring how the CDA signal might vary with memory load in the case in which study items are presented in sequential rather than simultaneous fashion.

Experiment

Methods

Subjects

15 Volunteers (20-36 years of age) participated in the experiment in exchange for monetary compensation. All participants had normal color vision, no history of neurological problems, and normal or corrected-to-normal vision acuity.

Stimuli and Apparatus

The stimuli were drawn from a pool of 2,400 unique object images obtained from the website of Talia Konkel and described by Bradly, Konkle, Alvarez, and Oliva (2008). Participants viewed the stimuli at a distance of 95 cm, displayed on a grey background with a 0.25 cm thick square that framed each image in either green (RGB value [0 255 0]) or red (RGB value [255 0 0]). The stimuli were presented on a Mac with Psychtoolbox (Brainard, 1997).

Procedure

Each subject completed two practice blocks (one in the VM condition and one in the CM condition) followed by four EEG recording blocks (two VM blocks and two CM blocks randomly ordered). The practice blocks were meant to familiarize subjects with the test and the CM vs. VM manipulations. Each practice block contained 50 trials and each EEG recording block contained 100 trials. In all conditions, half the test probes were targets and half foils.

For each block, 16 images were sampled without replacement. Subjects were tested on 8 of the images (stimulus-set) and the other 8 images served as filler images during study (filler-set). The filler images were never selected to serve as test probes. There were no overlapping images

between blocks. On each trial in the VM condition, a memory set of 2 or 4 items was randomly selected from the stimulus set and the items were presented sequentially for the subject to study. The presentation of the memory set was followed by the presentation of a test probe. Subjects indicated whether the test probe was "old" (a target item that was a member of the study list) or "new" (a foil item that was not a member of the study list) by left clicking or right clicking, accordingly. Test probes that were targets ("old") were randomly chosen from the memory set; test probes that were foils ("new") were randomly chosen from the remaining stimulus-set items that were not members of the memory set on the current trial. In the CM condition, 4 items from the stimulus set were randomly selected to serve as "target set" items and these stayed fixed across the block; the remaining items from the stimulus set became the "foil set" and these also stayed fixed. On each trial, a memory set of 2 or 4 items was always randomly selected from the target set. Just as in the VM condition, the items were presented sequentially for the subjects to study, and this study list was then followed by a test probe. Test probes that were targets ("old") were randomly chosen from the memory set; test probes that were foils ("new") were always randomly chosen from the fixed foil set.

A schematic illustration of a typical trial with set size two is presented in Figure 1. Subjects started each trial by clicking both keys of the mouse when a letter "B" was displayed at the center of the screen (visual angle of 0.2°). After a 500 ms delay, the memory set items were presented sequentially, each accompanied with a filler image that was randomly selected from the filler set. Each image was 10cm x 10cm in size. The memory set item and the filler image were simultaneously presented with one image on the right side and the other image on the left side of the fixation point (the inner border of each image was 5cm away from the fixation point; the visual angle to the center of the image is 6.37°). The images were distinguished with color

frames (red vs. green) and subjects were instructed to pay attention only to images framed by the task-relevant color (fixed across all blocks). In 50% of the trials, the study items stayed at the same side of the fixation point across the sequential presentation of the memory set; in the remaining 50% of trials the side of the study items was chosen randomly on each sequential presentation. In total, roughly 67% of trials were stay trials. The images were presented for 100ms followed by 900ms with just the fixation point. Following the presentation of the last memory set item, there was a 1000ms delay, after which a test probe was presented. The test probe (half the time a target) was presented at the center of the screen with the target-color frame. The test probe remained on the screen for 1.5 s or until the subject clicked the mouse key to make a response. Feedback was then provided with tunes in different pitch: high pitch indicated a correct response; low pitch indicated an incorrect response; a burst of three tunes indicated a slow response.

Prior to the practice blocks, subjects were informed of the task-relevant color (red or green, counterbalanced between subjects) and of the nature of the memory search task without information regarding the CM vs. VM manipulation. After completing the practice blocks, subjects were asked to verbally describe to the experimenter the difference between the 2 blocks and were informed about the CM vs. VM manipulation. After EEG net capping, each subject was asked to perform 6 eye-blink trials and 24 horizontal eye-movements (with 4 trials for each of 2.67, 5.15 and 10.29 degrees of eye-movements to the left or right of the center of the screen) before the start the memory search task.

Electroencephlogram acquisition and pre-processing

The electroencephalogram (EEG) was sampled at 32 channels at 1000hz and down sampled to 500hz. The signals were amplified by a factor of 20,000 using Sensorium amplifiers

with an analog bandpass filter of 0.01-100HZ. Eye-movements were monitored with electrodes 2 cm away from the eyes to capture horizontal eye-movements and an electrode was placed under the right eye to detect eye-blinks and vertical eye-movements. The data was later low-pass filtered below 50hz.

For each trial, the EEG data were collected 500ms prior to the onset of the first study item and 1500ms after the onset of the test probe. We used three steps to remove artifacts from the average ERP. The horizontal EOG from the instructed eye-movement trials were used to generate a linear function of degrees of eye-movement; we rejected trials with at least 4 degrees of horizontal eye-movement during the presentation of the memory set. In addition, two subjects were rejected for excessive eye-movement (>35% of trials). Research assistants in the lab also rejected any trials with obvious artifacts. EEGlab toolbox (Delorme & Makeig, 2004) was employed for EEG data analysis. For the 13 remaining subjects, an average of 14% trials were removed. Independent Component Analysis (ICA) was used to identify artifacts including eye-blinks, eye-movement, and muscle activity. The artifacts were subtracted from the raw EEG data prior to ERP analyses and Alpha power analyses. Due to the relatively low frequency of error trials (resulting in inadequate statistical power), we included only correct trials in the EEG analyses.

Behavioral Results

In Figure 2 we plot the probability of errors and the mean response time (RT) for correct trials as a joint function of condition (CM vs. VM), test-probe type (target vs. foil), and memory set size (2 vs. 4). The results are consistent with patterns observed in many previous studies of VM and CM memory search: RTs are much shorter and error rates are much lower in the CM

condition than in the VM condition. Most importantly, while VM error rates and RTs increased strongly with set size, CM performance stayed the same across set sizes. Such results indicate that the paradigmatic changes made in order to implement this EEG experiment did not alter the usual pattern of behavioral results.

To analyze the data, we applied a 2 (CM, VM) x 2 (Target, Foil) x 2 (set size 2, 4) repeated measure ANOVA to both the accuracy and RT data. For the accuracy data, the effects of both condition ($F(1,12)=12.25$, $p=0.004$) and set size ($F(1,12)=39.13$, $p<0.001$) were significant. The interaction between condition and set size ($F(1,12)=41.81$, $p<0.001$) was also significant, reflecting that set size had a big impact in the VM condition but not in the CM condition. For the RT data, the main effect of conditions was significant ($F(1,12)=10.29$, $p=0.008$). The interaction between condition and set size was marginally significant, $F(1,12)=3.58$, $p=0.083$, reflecting that set size again tended to have a bigger impact in the VM condition than in the CM condition.

EEG Analyses

CDA Analyses

In Figures 3A and 3B we show the average waveforms of lateral occipital-temporal electrodes (PO3/4, O1/2, PO7/8, P7/8), collapsed based on their relative locations to the stimuli during memory-set presentation (i.e., ipsilateral vs. contralateral). (Figure 3A shows the results for the set-size-4 trials, and Figure 3B for the set-size-2 trials.) To avoid any complications arising from conflicting CDAs due to swapping sides, we analyzed only those trials where the target stimuli stayed at the same side of fixation. The space between the contralateral waves and the ipsilateral waves measures the CDA. As shown in the figure, for both set sizes, the CDA is

observed in both the CM and VM conditions, although the magnitude of CDA is reduced in the CM condition compared to the VM condition. To bring out this result more clearly, in Figure 3C we plot the CDA in the CM and VM conditions for the first and second study items, averaged across the set-size-4 and set-size-2 conditions.

We performed a 2 (CM vs. VM) x 2 (Contralateral vs. Ipsilateral) x 2 (set size 2 vs. 4) repeated ANOVA of the averaged electrodes voltage during the 300-1000ms epoch after the onset of each study item. We found a significant main effect of relative sides (Contralateral vs. Ipsilateral, $F(1,12)=23.2$, $p<0.001$). Most important, the interaction between relative sides and condition was also significant ($F(1,12)= 6.63$, $p=0.024$), reflecting the reduced CDA in the CM condition compared to the VM condition.

Alpha Power Suppression

Researchers have shown that suppression of alpha power is associated with load in short-term memory (Fukuda & Woodman, 2017). Therefore, we decided to assess suppression of Alpha power in our study. EEG from parieto-occipital channels (P3/4, PO3/4, O1/2, Pz) of each trial was subjected to spectral decomposition using EEGLAB function "newtimef" with 3 cycles per morlet wavelet. We define the baseline as the mean Alpha power spectrum (8-13 HZ band) during the pre-trial time window (-500 to 0ms relative to the onset of the first study item). The percentage change of Alpha power for the memory set presentation relative to the baseline is then plotted in Figure 4. The average change of Alpha power is collapsed across electrodes from both sides of the scalp. (We also examined the Alpha power change separately for electrodes located contralateral vs. ipsilateral to the study items and found no difference in the pattern of results.) As shown in the figure, Alpha power reduced substantially after the onset of each study

item and there appears to be more reduction in the VM condition than in the CM condition. We averaged the change of Alpha power from baseline over the epoch of 300-1000ms after the onset of each study item. We performed a 2 (CM vs. VM) x 2 (set size 2 vs. 4) repeated ANOVA for the mean change of alpha power. The effect of condition was marginally significant ($F(1,12)=3.2$, $p=0.099$). None of the interactions were significant. Although the noise in these data makes any strong conclusions difficult, the results are consistent with those from the CDA analyses in showing greater Alpha power suppression in VM than in CM.

Effects of increasing the short-term memory load

The VM behavioral data show a decline in performance when load increases from two to four items to be remembered, a universal finding in the field. There is no hint, however, of an increase in the CDA as additional items are presented for study. This observation is supported by statistical test: A pairwise t-test (first vs. second study item) of average CDA over 300-1000ms after the onset of each study item revealed no evidence of a difference ($t<1$). There is also very little evidence for an increase in alpha power suppression as additional items are presented. Such findings suggest a refinement of the interpretation of the meaning of the CDA and alpha power suppression findings. We suggest they show load effects for the amount of information that an observer attempts to actively and simultaneously maintain in visual working memory. Under our conditions of testing, observers may have tried to actively maintain only the most recently presented item, without attempts to actively maintain the previous study items. Much future research will be needed to test this and numerous other possibilities.

Discussion

Limits on capacity of short term memories, defined by numbers of distinct items or by persistence, have been acknowledged and studied since the first days of psychology. Schneider and Shiffrin (1977) and Shiffrin and Schneider (1977) showed how consistent practice could overcome such limits through the development of automaticity, with a likely mechanism involving the retrieval from long-term memory of stored instances of the consistently mapped item-response pairs (e.g. Logan, 1988). Both these results are seen in the behavioral results from the present studies of probe-recognition memory search. The VM conditions show the effects of load or capacity limitations, with observers performing worse in cases in which four rather than two items are held in memory. This decline in performance was observed for both accuracy and response time measures. By contrast, as a result of consistent practice, the effects of memory load were greatly reduced in the CM conditions.

Recent years have seen the discovery of neural measurements that signify the presence of short-term memory load and capacity limitations. A few are based on EEG measures, including the CDA that was the focus of the present investigation (Carlisle et al., 2011; Luria et al., 2016; Vogel & Machizawa, 2004; Woodman et al., 2013). The CDA is correlated with the amount of material being held in at least one kind of short-term visual memory. Researchers have shown not only a dependence of the CDA upon the demands for memory maintenance, but also a reduction of the CDA in CM practice conditions in which a single stimulus was repeatedly mapped to the same response (Carlisle et al., 2011; Reinhart, Carlisle, & Woodman, 2014; Reinhart & Woodman, 2014). Recent work has also indicated that the magnitude of alpha-band suppression can provide a reliable neural metric of storage in visual working memory (e.g.,

Fukuda, Kang, & Woodman, 2016; Fukuda & Woodman, 2017); thus, we also quantified this activity.

Here we measured EEG while subjects were sequentially shown a substantial number of complex pictures, and with considerable training in both VM and CM, deviating from previous work in these regards. Both the CDA and the amount of alpha power suppression were measured after each presentation of the study items. Both the magnitude of CDA and alpha suppression were greater for VM than CM. These results were consistent with the hypothesis derived from behavioral and formal modeling work that practice under the present kinds of CM conditions did indeed reduce the demands for short term memory capacity. Furthermore, previous demonstrations of the reduced CDA under CM conditions involved the repetition of only a single target item across consecutive trials. Our results generalize that finding by showing a reduced CDA under CM conditions involving large sets of to-be-remembered stimuli and in which the test probes are spaced throughout the entire training block.

One other finding, however, was not expected a priori: In VM, as additional pictures were presented sequentially, the size of the CDA and the amount of alpha suppression did not increase, despite the behavioral evidence that load in short-term memory was increasing. As noted earlier, a number of studies using CDA have shown that an increase in memory load increases the CDA. There are several possible explanations for the difference in findings between the present experiment and previous studies of the CDA. One possibility is that the CDA measures the load associated with attempts to actively maintain multiple items, whereas in our probe-recognition experiments the subjects may have tried to actively maintain only the most recently presented item. We plan to pursue this and other possibilities in future research.

References

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, USA, 105,* 14325–14329.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436.

Cao, R., Shiffrin, R. M., & Nosofsky, R. M. (2018). Item frequency in probe-recognition memory search: Converging evidence for a role of item-response learning. *Memory & cognition,46,* 450-463.

Carlisle, N. B., Arita, J. T., Pardo, D., & Woodman, G. F. (2011). Attentional templates in visual working memory. *Journal of Neuroscience, 31, 9315–9322*.

Delorme, A. & Makeig, A. (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics. *Journal of Neuroscience Methods 134*:9-21.

Fukuda, K., Kang, M.-K., & Woodman, G. F. (2016). Distinct neural mechanisms for spatially lateralized and spatially global working memory representations. *Journal of Neurophysiology, 116,* 1715-1727.

Fukuda, K., & Woodman, G. F. (2017). Working memory buffers information retrieved from human long-term memory. *Proceedings of the National Academy of Sciences, 114*(20), 5306-5311.

Logan, G.D. (1988). Toward an instance theory of automatization. *Psychological Review, 95, 492-527*.

Luria, R., Balaban, H., Awh, E., & Vogel, E. K. (2016). The contralateral delay activity as a neural measure of visual working memory. *Neuroscience & Behavioral Reviews, 62,* 100 108.

Nosofsky, R.M., Cao, R., Cox, G.E., & Shiffrin, R.M. (2014). Familiarity and categorization processes in memory search. Cognitive Psychology, 75, 97-129.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104(2), 266-300.*

Reinhart, R. M. G., Carlisle, N. B., & Woodman, G. F. (2014). Visual working memory gives up attentional control early in learning: Ruling out inter-hemispheric cancellation. *Psychophysiology, 51*(800-804).

Reinhart, R. M. G., & Woodman, G. F. (2014). High stakes trigger the use of multiple memories to enhance the control of attention. *Cerebral Cortex, 24,* 2022-2035.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84,* 1-66.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84,* 127-190.

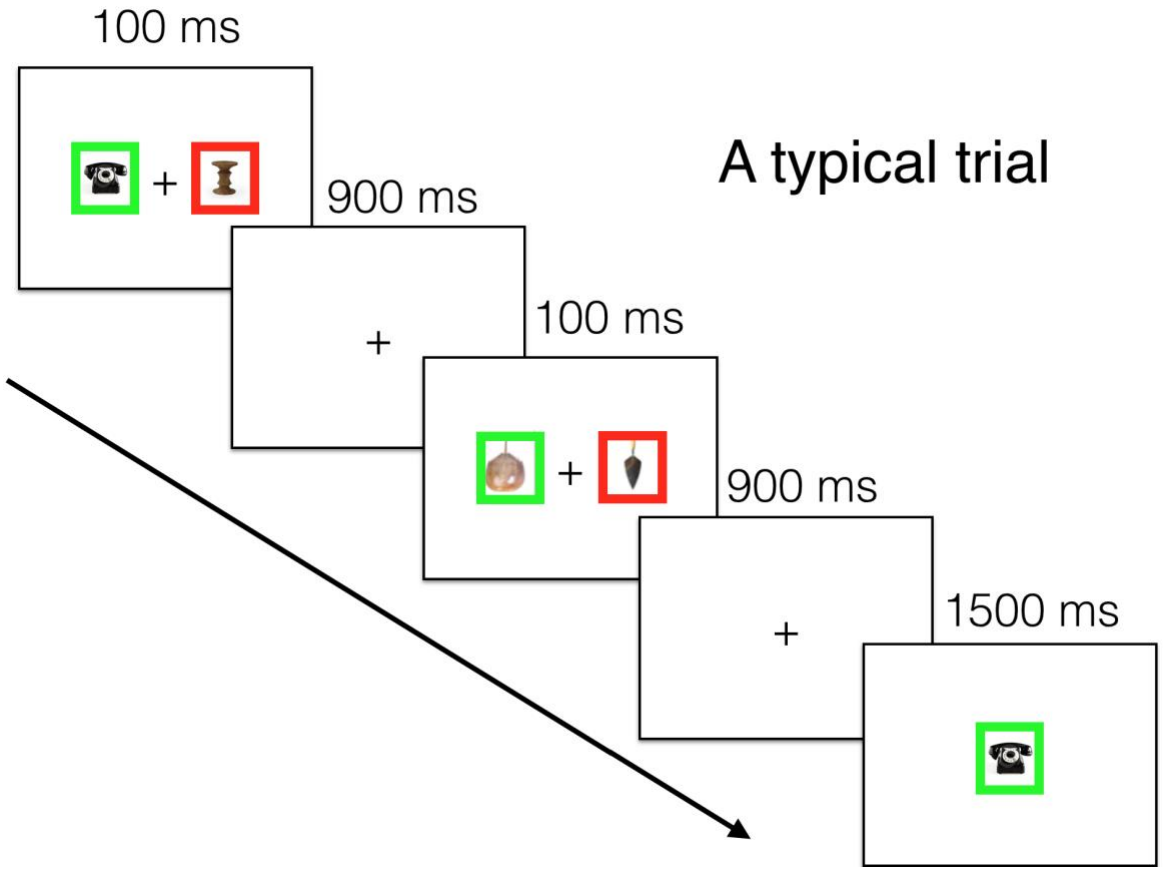
Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied, 74*(11), 1-29.

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature, 428(6984)*, 748-751.

Woodman, G. F., Carlisle, N. B., & Reinhart, R. M. (2013). Where do we store the memory representations that guide attention?. *Journal of Vision, 13(3), 1-1*.

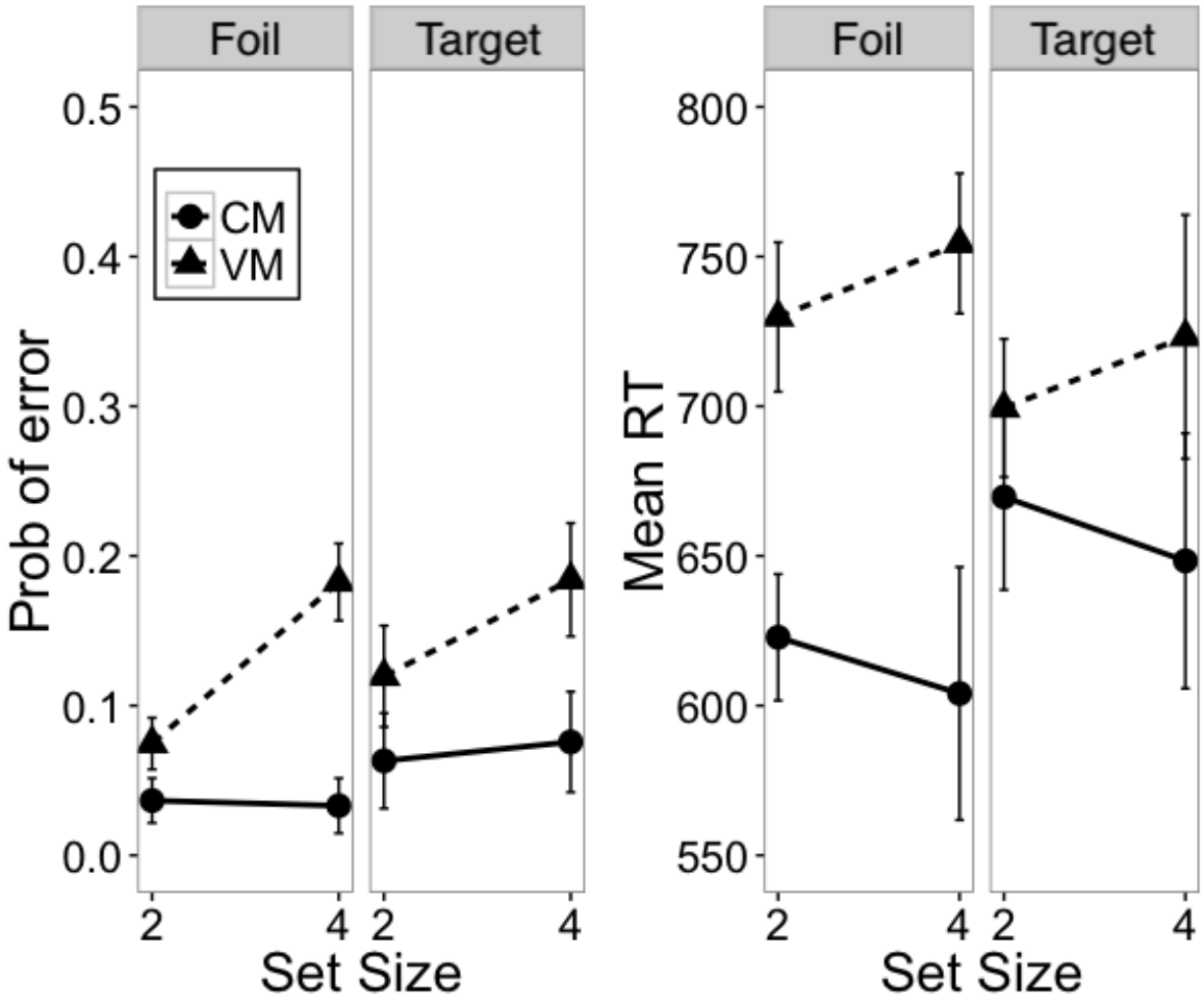
Figures

Figure 1



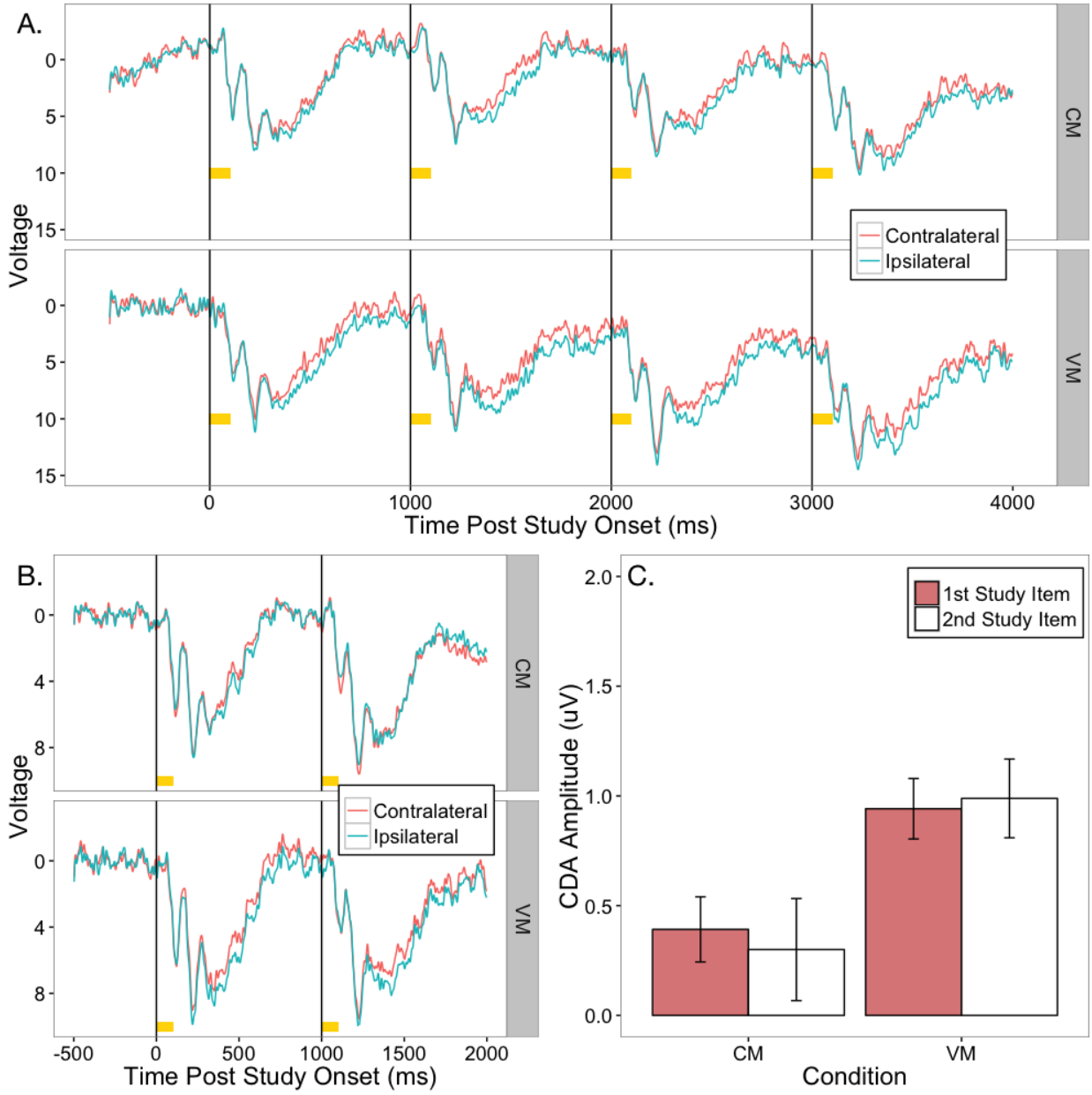
Example of one trial in the experiment (set size 2).

Figure 2



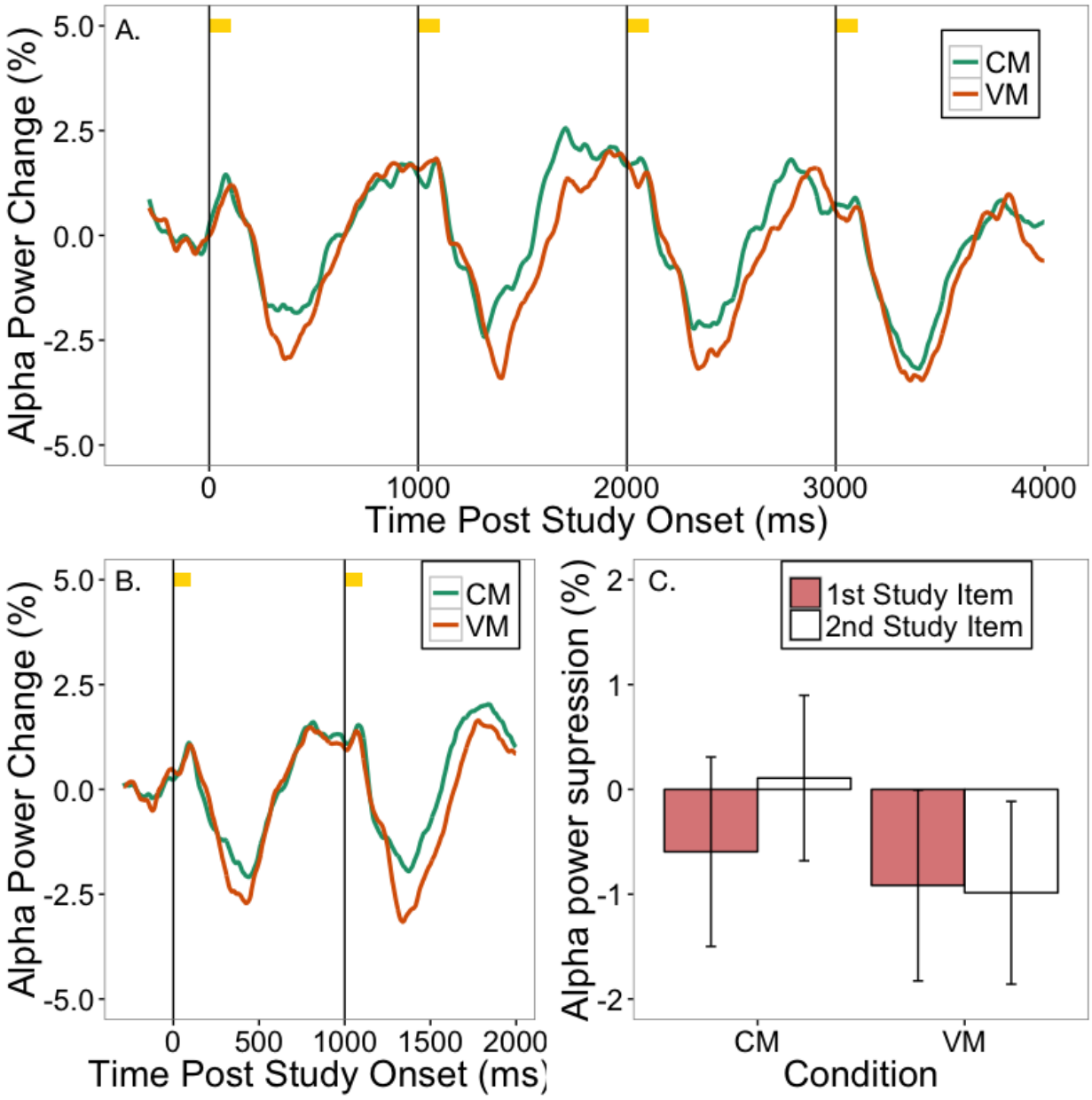
Probability of error (left panel) and mean correct response time (right panel) as function of condition (CM, VM), set size (2, 4) and probe type (foil, target)

Figure 3



CDA signals in the experiment. A. Grand average waveforms from lateral occipital-temporal electrodes for set size 4. The vertical black lines indicate onset of each study item and the yellow shades indicate the duration of study item presentations (same applies to B). B. Grand average waveforms from lateral occipital-temporal electrodes for set size 2. C. Grand average of CDA in the time period 300-1000 ms post the onset of the first and second study item.

Figure 4.



Alpha-power change during study. A. Grand average waveforms from lateral occipital-temporal electrodes for set size 4. The vertical black lines indicate onset of each study item and the yellow shades indicate the duration of study item presentations (same applies to B).

B. Grand average waveforms from lateral occipital-temporal electrodes for set size 2. C. Grand average of alpha power suppression in the time periods 300-1000 ms post the onset of the first and second study item

Rui Cao

caorui.beilia@gmail.com

Education

Indiana University | 2012 August–present
PhD program, Cognitive Psychology with Minor in Computational Neural Modeling
Co-Advisors: Richard Shiffrin and Robert Nosofsky

The Ohio State University | 2008–2012 June
B. S., Psychology with Research Distinction
B. S., Actuarial Science with Minor in Statistics
Advisor: Simon Dennis

Publication

Cao, R., Busey, T.A., Nosofsky, R. M., Shiffrin, R. M., & Woodman, G.F. (2018).
Tracking the Development of Automaticity in Memory Search with Human Electrophysiology
40th Annual Cognitive Science Society Meeting Proceedings

Cao, R., Shiffrin, R. M., & Nosofsky, R. M. (2018).
Item Frequency in Probe-Recognition Memory Search: Converging Evidence for a Role of Item-Response Learning
Memory & Cognition

Cao, R., Nosofsky, R. M., & Shiffrin, R. M. (2016).
The development of Automaticity in Short-Term Memory Search: Item-Response Learning and Category Learning
Journal of Experiment Psychology: Learning, Memory, and Cognition

Nosofsky, R. M., **Cao, R**., Cox, G. E., & Shiffrin, R. M. (2014).
Familiarity and categorization processes in memory search.
Cognitive psychology

Nosofsky, R. M., Cox, G. E., **Cao, R.**, & Shiffrin, R. M. (2014).
An exemplar–familiarity model predicts short-term and long-term probe recognition across diverse forms of memory search.
Journal of Experimental Psychology: Learning, Memory, and Cognition

Selected Academic Presentation

02/2017 "Item-Response Learning in Memory Search"
Australian Mathematical Psychology Conference 2017

08/2016 "Learning to Search Short-Term Memory: Item or Category"
2016 Annual Meeting of Society of the Cognitive Science Society

08/2016 "Item Response Learning vs. Familiarity process in Memory Search"
2016 Annual Meeting of Society of Mathematic Psychology

04/2016 "Item Learning vs. Category Learning in Memory Search"
Invited talk at Xi'an Jiaotong University, China

11/2015 "Item Learning vs. High-Level Categorization in Consistent-Mapping Memory Search"
56th Annual Meeting of the Psychonomic Society

11/2014 "The Categorization and Familiarity Processes in Memory Search"
Psychonomics Society's 55 Annual Meeting

07/2013 "The Dynamics of Intrusion in Cued Recall"
2013 Annual Meeting of Society of Mathematic Psychology

05/2013 "Target or Foil: Which Causes Output Interference?"
The 2013 Context and Episodic Memory Symposium

05/2012 "Word Frequency and List Length Effect on Cued Recall"
2012 Annual Meeting of Society of Mathematic Psychology

Organization & Teaching Experience

Indiana University Bloomington | Fall 2013–Spring 2015
Cognitive Lunch Colloquium Organizer

2012 Annual Meeting of Society of Mathematical Psychology | May 2012
Organizing Volunteer

Experimental Methods in Psychology in Indiana University | Fall 2015
Lab Instructor

Statistical Techniques in Indiana University | Fall 2016, Summer 2017
Assistant Instructor

Human Learning and Cognition in Indiana University | Fall 2013–Spring 2014
Assistant Instructor

Cognitive Psychology in Indiana University | Fall 2012–Spring 2013
Assistant Instructor

Mathematics and Statistics Learning Center in Ohio State University | Fall 2013–Spring 2015
Student Tutor

Scholarships & Travel Funding

2017-2018	Dissertation Year Fellowship
May 2015	NSF Sackler Colloquium- Drawing Casual Inference from Big Data
2013-2017	Provost's Award for Women in Science (Won multiple times)
April 2014	Student Travel Award for Context & Episodic Memory Symposium
Summer 2011	Psychology-Undergraduate Research Office Fellowship
May 2011	Psychology Conference Travel Scholarship
2008–2012	International Undergraduate Scholarship